UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Friday 14th May 2010

14:30 to 16:30

Year 3 Courses

Convener: K. Kalorloti
External Examiners: K. Eder, A. Frisch

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and TWO other questions.

Question 1 is COMPULSORY.

The short answer parts of Question 1 are each worth 1 mark, 20 marks in total. Each of the remaining three questions is worth 15 marks — answer any TWO of these.

Use one script book for each question, that is, three books in all.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

Part A

Answer ALL questions in Part A.

Your answers should be thorough—take the opportunity to show what you know, as opposed to being as brief as possible—but they need not be lengthy. When two terms are contrasted, make sure your short definitions of each make clear where the contrast lies. Each question is worth one mark, 20 marks in total for this section.

1. What does the most common frequency distribution of natural language items look like? What is its name?
2. What is a **gold standard** in data-intensive linguistics?
3. Which kind of statistic, **parametric** or **non-parametric**, is usually required for evaluating natural language processing systems? Why?
4. What's the difference between a **representative** corpus and an **opportunistic** one?
5. What is corpus **metadata**? Give four examples.
6. What are some of the reasons for the rise of **data-intensive** approaches to speech and language processing?
7. What is the difference between **well-formed** XML and **valid** XML?
8. Give an equation relating **joint** and **conditional** probabilities. What is the difference between the two?
9. What is an **N-gram language model**?
10. What is the rule of thumb for estimating how much data is required to give a good N-gram model? How much data does this suggest we need for a good tri-letter model of English words?
11. What is **backoff** and why is it necessary?
12. What is **smoothing** in the context of language models? Name two examples.
13. Give the formula for **Bayes' rule**.
14. In the noisy channel model, what are the **likelihood** and the **prior**?
15. What is the difference between a **Markov chain** and a **Hidden Markov Model**?

16. What are the three main computations associated with **Hidden Markov Models**?

17. What are the two main sources of **ambiguity** in the parsing of natural language text?

18. What is a **well-formed substring table** and what use is it?

19. What does it mean to **lexicalise** a **context-free phrase-structure grammar**?

20. Illustrate the two most common forms of **category** in a **categorial grammar** and explain what they stand for.

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 is it? And if so, which of the famous sisters wrote it?

   Drawing on the language modelling technologies discussed in lectures and tutorials, 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) Describe in detail the steps you would go through to answer the two questions. Be sure to set out your background assumptions and the hypotheses you would be trying to test. Also, be sure to describe the modelling technique(s) you would use, how you would train the models and how you would use them to confirm or deny your hypotheses. [10 marks]

   (b) What reasons are there to be suspicious of the outcome of your experiments? Which result would you have more confidence in (as between "yes, the manuscript is the work of Xxx Brontë" and "no, not the work of either sister"), and why? [5 marks]

2. Dynamic programming and Hidden Markov Models

(a) Both Jurafsky & Martin and the course notes gave worked examples of using dynamic programming to find the minimum edit distance between two strings.

which is cheaper to turn into `'gold'`, `'lead'` or `'cola'`, given the standard costs of 1 for both deletion and insertion and 2 for substitution? [2 marks]

(b) Describe how **dynamic programming** can be used to compute minimum edit distance. Include a diagram showing, in full detail, one step in the computation of the cost of the correction of `'lead'` to `'gold'`. [4 marks]

(c) Using the language and channel models for an HMM from **Appendix A**, diagram the computation of the total probability of the French output "les bonnes dorment". You should draw a sequence of lattices, one for each HMM transition.

Don't forget both the initial (from sentence start) and final (to sentence end) transitions.

Show your work. That is, make clear *how* you arrive at the probability you enter in each cell of the lattice. [9 marks]

3. Best-first chart parsing

   Using the probabilistic CF-PSG from **Appendix B** simulate a best-first top-down chart parser parsing the sentence "time flies".

   Use the technique from Jurafsky & Martin, also used in lectures, of drawing actual edges in your chart for inactive edges and partially complete active edges, but just writing the dotted rule for empty active edges below the relevant vertex.

   Show both the chart and the agenda, crossing items off the agenda as you draw them into the chart, and *numbering* both the agenda entry and the chart entry as you do so to show the order in which things were done.

   Start with the following two edges already in the chart:

   ${}_0\text{time}_1$
   ${}_1\text{flies}_2$

   with both edges having 0 cost.

   Start with the following entry in the agenda:

   ${}_0Top \longrightarrow \cdot S_0$

   with 0 cost and the best possible figure of merit.

   Be sure to show the cost of every edge in both chart and agenda, and the figure of merit for edges in the agenda.

   Mark each edge in the agenda other than the first (top-down initiator) one as either TD (top-down) or F (fundamental), to show which rule they were 'built' by.

   Explain how you are calculating costs and the figure of merit you are using, and why. [15 marks]

Appendix A: Language and channel models

These models are for use in answering Question 2 in Part B.

Language model

	the	them	good	maids	sleep	$
.	0.12	0.00002	0.0008	0.00001	0.00003	0.0
the	0.00008	0.00001	0.0006	0.00001	0.00001	0.0001
them	0.018	0.00002	0.0004	0.0	0.00007	0.2
good	0.005	0.0	0.0008	0.0	0.0002	0.063
maids	0.01	0.0	0.0	0.0	0.044	0.11
sleep	0.009	0.00016	0.00032	0.0	0.0008	0.25

Channel model

	the	them	good	maids	sleep
les	0.2	0.47	0.0	0.0	0.0
le	0.23	0.0	0.0	0.0	0.0
la	0.31	0.0	0.0	0.0	0.0
eux	0.0	0.38	0.0	0.0	0.0
elles	0.0	0.16	0.0	0.0	0.0
bonnes	0.0	0.0	0.063	0.018	0.0
bonne	0.0	0.0	0.28	0.0	0.0
honnête	0.0	0.0	0.01	0.0	0.0
servantes	0.0	0.0	0.0	0.1	0.0
dormir	0.0	0.0	0.0	0.0	0.29
dormez	0.0	0.0	0.0	0.0	0.008
dorment	0.0	0.0	0.0	0.0	0.13

Note: read these tables as follows: The language model has start states down the left and destination states across the top, so for example the probability of a transition from **the** *to* **good** *is* `0.0006`*; the channel model has English across the top and French down the left, so for example the probability of seeing the word* **bonnes** *in state* **maids** *is* `0.018`*. In the language model, full stop ('.') is used for sentence start and dollar-sign ('$') for sentence end.*

The language model probabilities were taken from the BNC: the unexpectedly high probability for the bigram "the the" is due almost entirely to errors in the texts. Whether those errors are in the originals, or were introduced by the transcription process, is not clear.

Appendix B: Probabilistic CF-PSG

These rules are for use in answering Question 3 in Part B.

Note: the numbers given are costs

Rule	Cost
S $\longrightarrow$ NP VP	2.3
S $\longrightarrow$ VP	.3
VP $\longrightarrow$ V0	1.6
VP $\longrightarrow$ V1 NP	.6
V1 $\longrightarrow$ *time*	15
V0 $\longrightarrow$ *flies*	11
NP $\longrightarrow$ *time*	16.5
NP $\longrightarrow$ *flies*	12.5