UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Tuesday 17th August 2010

09:30 to 11:30

Year 3 Courses

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

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 2 marks, 20 marks in total. Each of the three questions in part B 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—as opposed to being as brief as possible—but they need not be lengthy: three or four sentences should do it. When two terms are contrasted, make sure your short definitions of each make clear where the contrast lies. Each question is worth two marks, 20 marks in total for this section.

1. What name is used for information *about* a corpus, as opposed to its language data contents? What kind of information do we expect to find?
2. What is the difference between **joint** and **conditional** probabilities? Give an equation relating the two.
3. What is the difference between **unigram**, **bigram** and **trigram** language models? What is the rule of thumb for estimating how much data is required to give a good N-gram model?
4. What problem arises from **sparse data** when constructing and using language models? What can be done about it?
5. What is **mutual information**? Give an example of a language processing task where it might be used.
6. Name and describe the two places where probabilities appear in a **Hidden Markov Model**?
7. What does the most common **frequency distribution** of natural language items look like? What is its name?
8. What is **POS-tagging**? Name one technology which can be used to perform it.
9. What is the relation of **cost** to **probability** in probabilistic speech and language processing? What advantage does **cost** offer?
10. What is the **noisy channel model**? What are its two main constituents?

Part B

Answer two questions in Part B.

1. 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.

   What is the difference in the edit distance for turning `'Mary'` into `'Mairi'` versus turning `'Vari'` into `'Mairi'`, 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 change from `'Mary'` to `'Mairi'`. [4 marks]

   (c) Viterbi search is a form of **dynamic programming** that finds the most likely path through an HMM for a given set of inputs.

   Using the language and channel models for an HMM from **Appendix A**, for a system recovering the English translation of French, draw a detailed digram of one cell in the computation for the cost of the most likely translation for "les bonnes dorment", namely the one for 'maids' being the translation of 'bonnes'. For the previous column, use 5.4 as the lowest cost for 'the' being the translation of 'les', 16.9 as the lowest cost for 'them' being the translation of 'les', and infinity as the cost for the other three possible translations of 'les'.

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

2. Determining text authorship

The *Federalist Papers* are a collection of 85 anonymously-authored political articles written in 18th century America. They were in fact mostly authored by Alexander Hamilton and James Madison, but the historical evidence for identifying the author of the individual articles is not always definitive.

We can divide the articles into three categories:

- Those asserted to be by Hamilton;
- Those asserted to be by Madison;
- Those whose authorship has not been established.

We'd like to know which assertions are correct and which are incorrect, and for the incorrect and unknown cases, whether the likely author is Hamilton, Madison or a third party.

Drawing on the language modelling technologies discussed in lectures and tutorials, design an experiment to answer these questions, assuming you have access to electronic versions of all 85 articles, as well as substantial amounts of representative 18th-century American political writing by a range of authors.

(a) Describe in detail the steps you would go through to answer the 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 factors will determine the reliability of your results? In general, which is likely to be more reliable: "these are similar" or "these are different", in this sort of experiment? [5 marks]

3. Chart parsing

Outline the algoritm for best-first bottom-up (left-corner) chart parsing using probabilistic context-free phrase structure grammars.

Include descriptions of the chart and its constituent parts, the agenda and the grammar.

Describe the parsing algorithm from initiation through normal operation to conclusion, describing how processing rules determine what operations are performed on the chart and agenda. Use costs throughout, rather than probabilities as such.

Use the probabilistic CF-PSG from **Appendix B** to illustrate the key operations of the algorithm. 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. In each illustration, include only the relevant parts of the chart and agenda.

[15 marks]

Appendix A: Language and channel models

These models are for use in answering Question 1 in Part B.

All numbers are *costs*, that is, negative log probilities, so you can a) sum them, rather than multiplying and b) you are looking to *minimise* the total cost.

Language model

	the	them	good	maids	sleep	$
.	3.1	15.8	10.3	20.2	14.9	∞
the	13.7	19.0	10.8	16.6	16.2	13.2
them	5.8	15.4	11.4	∞	13.9	2.3
good	7.7	∞	10.3	∞	12.1	4.0
maids	6.7	∞	∞	∞	4.5	3.2
sleep	6.8	12.6	11.6	∞	10.3	2.0

Channel model

	the	them	good	maids	sleep
les	2.3	1.1	∞	∞	∞
le	2.1	∞	∞	∞	∞
la	1.7	∞	∞	∞	∞
eux	∞	1.4	∞	∞	∞
elles	∞	2.6	∞	∞	∞
bonnes	∞	∞	4.0	5.8	∞
bonne	∞	∞	1.9	∞	∞
honnête	∞	∞	6.6	∞	∞
servantes	∞	∞	∞	3.3	∞
dormir	∞	∞	∞	∞	1.8
dormez	∞	∞	∞	∞	6.9
dorment	∞	∞	∞	∞	3.0

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 cost of a transition from **the** *to* **good** *is* `10.8`*; the channel model has English across the top and French down the left, so for example the cost of seeing the word 'honnête' in state* **good** *is* `6.6`*. In the language model, full stop ('.') is used for sentence start and dollar-sign ('$') for sentence end.*

The language model cost were taken from the BNC: the unexpectedly low cost 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