UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Thursday 14 May 2009

14:30 to 16:30

Year 3 Courses

Convener: K Kalorkoti
External Examiners: A Frisch, J Gurd

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—you MUST answer exactly TWO of these.

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

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

1. Short answer questions—**answer** ALL **sub-questions**

Your answers should be more than a single sentence—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.

(a) What's the difference between a **selective** account and an **instructional** account of, for example, speech recognition? [1 mark]

(b) What is **Zipf's law**? [1 mark]

(c) Contrast the **inherent** and **analytical** properties which you might find annotated in a corpus. [1 mark]

(d) What's the difference between a **representative** corpus and an **opportunistic** one? [1 mark]

(e) What is corpus **metadata**? Give four examples. [1 mark]

(f) What are two of the benefits of using a markup language (SGML, XML) to deliver corpora? [1 mark]

(g) What is the difference between **well-formed** XML and **valid** XML? [1 mark]

(h) What is a **significance test** and how is it used in evaluation? [1 mark]

(i) What is **conditional** probability? Give an equation defining it in terms of **joint** probability. [1 mark]

(j) What is an **N-gram language model**? [1 mark]

(k) What is **smoothing** in the context of language models? Name two examples. [1 mark]

(l) Give the formula for **Bayes' rule**. [1 mark]

(m) In the noisy channel model, what does **likelihood** refer to? [1 mark]

(n) What are two of the three main computations which need to be defined for **Hidden Markov Models**? [1 mark]

(o) What are the two main sources of **ambiguity** in the parsing of natural language text? [1 mark]

(p) What is the **fundamental rule** of chart parsing? [1 mark]

(q) What is the difference between **top-down** and **left-corner** invocation in chart parsing? [1 mark]

(r) What does it mean to **lexicalise** a **context-free phrase-structure grammar**? [1 mark]

(s) How do the **categories** of a **categorial grammar** differ from those of a **context-free phrase-structure grammar** [1 mark]

(t) What is **precision** and how is it used? [1 mark]

Answer two from among questions 2–4.

2. Exploring text corpora

Use the two corpus extracts in **Appendix A** to illustrate your points in answering this question.

Describe the operations you would perform in quantifying the properties of a corpus such as the Brown corpus, which is composed of material from qualitatively different sources. Frame your description in terms of exploring ways in which you might look for contrasts between the different samples in the corpus. Try to cover all of the following topics, recalling the exercises we have done in tutorials and assessed coursework:

- letter N-grams
- tokenisation
- word length
- word normalisation
- what constitutes a word
- frequency distribution
- types vs. tokens
- probability distribution
- plots
- sentence length
- word N-grams

Whereever possible, use either pseudocode or Python/NLTK to make your description precise. Minor errors in Python syntax or NLTK class names will *not* be penalised. [15 marks]

3. Dynamic programming and POS-tagging

Data for use in answering this question can be found in **Appendix B**.

(a) Both Jurafsky & Martin and the course notes gave worked examples of using dynamic programming to find the minimum edit distance between two strings. They both used naive costings for **substitutions**.

Suppose we make substitutions less expensive for letters which represent similar sounds. Given the following letter groups:

- aeiouy
- bp
- ckq
- dt
- lr
- mn
- hw
- gj
- fpv
- szc
- szx

we will say that substitutions *within* a group cost only 1.5, whereas substitutions that are not within *any* group cost 2. So for example substituting `f` for `v` will cost 1.5, but substituting `f` for g will cost 2.

Note that some letters occur in more than one group!

Given the standard deletion and insertion costs of 1 as well, draw a dynamic programming lattice to find the lowest-cost transformation of `catsup` into `ketchup`. Be sure to include backpointers.

Show your work. That is, make clear *how* you arrive at the cost you enter in each cell of the lattice. [6 marks]

(b) Using the language and channel models from **Appendix B**, diagram the Viterbi search for the most likely POS tag sequence for the sentence "the good sheep". All numbers are *costs*, that is, negative log probabilities, so you can a) sum them, rather than multiplying and b) you are looking to *minimise* the total cost.

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

---

4. Best-first chart parsing

   Using the probabilistic CF-PSG from **Appendix C** simulate a best-first left-corner chart parser parsing the sentence "time flies".

   Use the technique from Jurafasky & 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 agenda:

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

   with both entries having 0 cost and the best possible figure of merit. Break the resulting tie in favour of the 'flies' edge, i.e. add it to the chart first.

   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 two (lexical) ones as either LC (left-corner) 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: Sample corpus data

These extracts from the Brown corpus are for use in answering Question 2. The paragraphs below correspond to single lines in the originals.

Extract 1

```
Hemphill said that the Hughes Steel Erection Co. contracted to do the
work at an impossibly low cost with a bid that was far less than the
" legitimate " bids of competing contractors .

The Hughes concern then took " shortcuts " on the project but got
paid anyway , Hemphill said .

The Controller's charge of rigging was the latest development in an
investigation which also brought these disclosures Tuesday :

The city has sued for the full amount of the $172,400 performance bond
covering the contract .

The Philadelphia Transportation Co. is investigating the part its
organization played in reviewing the project .
```

Extract 2

```
Such a little thing to start with -- the car registration .

" Ida , where is the car license " ?  She asked .

" I can't find it in the glove compartment " .

" Vera must have it " , I answered readily enough , recalling her
last visit .

" Vera " , she was frowning .

" Why should Vera have it " ?

Had she forgotten she had signed the car away , that whatever they
mutually owned had been divided among the children ?

I was silent .

I didn't want to stir things up .
```

Appendix B: Language and channel models

These models are for use in answering Question 3.

Language model

	jj	at	nn	nns	$
.	2.3	0.7	3.3	3.3	∞
jj	2.3	∞	1.3	1.7	3.3
at	1.7	∞	1.3	1.7	∞
nn	∞	∞	3.3	3.3	0.3
nns	∞	∞	3.3	3.3	0.3

Channel model

	jj	at	nn	nns
the	∞	0.0	∞	∞
good	0.2	∞	2.3	∞
sheep	4.3	∞	2.3	2.3
man	4.3	∞	0.7	∞
men	∞	∞	∞	0.3

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 **at** *to* **nns** *is* `1.7`*; the channel model has parts of speech across the top and words down the left, so for example the cost of seeing the word 'man' in state* **nn** *is* `0.7`*. In the language model, full stop ('.') is used for sentence start and dollar-sign ('$') for sentence end.*

Appendix C: Probabilistic CF-PSG

These rules are for use in answering Question 4.

Note: the numbers given are costs

Rule	Cost
S $\longrightarrow$ NP VP	.57
S $\longrightarrow$ VP	6.8
VP $\longrightarrow$ V0	1.6
VP $\longrightarrow$ V1 NP	.6
V1 $\longrightarrow$ *time*	15
V0 $\longrightarrow$ *flies*	11
NP $\longrightarrow$ *time*	5.2
NP $\longrightarrow$ *flies*	13