UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Saturday 21st May 2011

14:30 to 16:30

Year 3 Courses

Convener: K. Kalorkoti
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 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 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: from three or four 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. What is a **frequency distribution**? What does the most common frequency distribution of natural language items look like? What is its name? Give two examples of items that have this distribution.
2. Name four applications for which language n-gram modelling is useful.
3. Assuming a bigram language model with $\langle s \rangle$ and $\langle /s \rangle$ respectively marking the beginning and end of a sentence, give an expression for the joint probability of the sentence "I want to meet my friends" made up of simpler probabilities.
4. In the noisy channel model, what are the **likelihood** (or **sensor model**) and the **prior** (or **transition model**)? Use these two to explain the difference between a **Markov chain** and a **Hidden Markov Model**.
5. Why is there no absolute measure for evaluating n-gram models? Define the *relative* measure people use instead for the intrinsic evaluation of n-gram models.
6. Define **backoff estimation** for n-gram models, and describe one potential advantage of backoff estimation over Good-Turing smoothing.
7. What does it mean to add **features** to a **context-free phrase structure grammar**? Give two examples of language phenomena which motivate this combination.
8. What are **selectional restrictions**? Consider the following example sentence:

   Cambridge voted conservative.

   In view of your answer about selectional restrictions, give an account of this sentence in terms of the word sense ambiguity for place names.

Part B

Answer two questions in Part B.

1. **Spell Checking**

   Suppose that you are developing a spell checker. Your aim is two-fold. First, you need to build a (binary) classifier, which maps each character string in the text that occurs between two spaces to either *correct* (i.e., the character string is spelled correctly) or *incorrect*. Secondly, you need to build a probabilistic model which maps each character string that is classified *incorrect* by the binary classifier to a ranked set of alternative correct spellings.

   (a) Minimum edit distance will clearly be an important source of information for estimating the ranked set of alternative correct spellings for an incorrectly spelled word. Assume that it costs 1 to delete or insert a character in a string, 2 to substitute one character for another, and 0 to leave a character unchanged.

   Draw a dynamic programming lattice to find the lowest-cost transformation from `form` to `from`. Show your work. Include the backtraces, making clear *how* you arrive at the cost you enter in each cell of the lattice. [*4 marks*]

   (b) Describe the kinds of contextual features and language resources that would potentially be informative for building the binary classifier for identifying incorrectly spelled words, and also those that would be informative for building a probabilistic model for ranking candidate corrections. Use illustrative examples, including examples of the character string and examples of the textual context in which it occurs, to justify your answer. [*9 marks*]

2. **Lexical Semantics**

(a) Explain the difference between homonymy and polysemy. Illustrate your answer with examples of each kind of ambiguity. [*3 marks*]

(b) According to WordNet, the word **scrap** has the following 4 noun senses and 3 verb senses:

Nouns:

sn1: bit, chip, flake, fleck, scrap
(a small fragment of something broken off from the whole)
a bit of rock caught him in the eye

sn2: rubbish, trash, scrap
(worthless material that is to be disposed of)

sn3: scrap
(a small piece of something that is left over after the rest has been used)
she jotted it on a scrap of paper; *there was not a scrap left*

sn4: fight, fighting, combat, scrap
(the act of fighting; any contest or struggle) *a fight broke out at the hockey game*;
there was fighting in the streets; *the unhappy couple got into a terrible scrap*

Verbs:

sv1: trash, junk, scrap
(dispose of (something useless or old))
trash these old chairs; *junk an old car*; *scrap your old computer*

sv2: quarrel, dispute, scrap, argufy, altercate
(have a disagreement over something)
We quarrelled over the question as to who discovered America; *These two fellows are always scrapping over something*

sv3: scrap
(make into scrap or refuse) *scrap the old aeroplane and sell the parts*

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. [*4 marks*]

(c) Explain why polysemous sense ambiguity creates a major challenge for creating an online lexical resource that defines the meaning of words, such as WordNet. Justify your answer by using examples. [*6 marks*]

3. **Hidden Markov Models**

Given a **Hidden Markov Model**, Viterbi search can be used to determine the most probable state sequence s_1^n for any given sequence of observations o_1^n, that is

$$\operatorname*{argmax}_{s_1^n} P(s_1^n|o_1^n)$$

(a) Give the formula for Bayes' Rule, and use it to transform the $P(...)$ expression above into its more useful form consisting of the product of two probabilities. Identify which of these is the **prior** and which the **likelihood**. [*4 marks*]

(b) HMMs can be used to translate from one language to another, with the input observations being the source language words, and the hidden state labels being the target language words.

In Appendix A you will find language and channel models for translating French into English. Below is a partially-completed diagram of the Viterbi search for finding the cost of the lowest-cost English translation for the sentence "les bonnes dorment", based on those models. 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.

the	5.4	∞	∞
them	16.7	∞	∞
good	∞	20.1	∞
maids	∞	27.8	∞
sleep	∞	∞	
	les	bonnes	dorment

i. Explain how the value `5.4` in the upper-left-hand cell is computed. [*2 marks*]

ii. Compute the correct value for the empty lower-right-hand cell. Show your work. That is, make clear *how* you arrive at your result. [*3 marks*]

iii. If we had wanted the *total probability* of the translation, how would the nature of the computation for each cell have to change? Do *not* recompute the lattice, just describe how you would do so. Include an explanation of why total *cost* is expensive to compute, as opposed to total *probability*. [*4 marks*]

Appendix A: Language and channel models

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

		the	them	good	maids	sleep	$
Language model	.	3.1	15.6	10.3	16.6	15	∞
	the	13.6	16.6	10.7	16.6	16.6	13.3
	them	5.8	15.6	11.3	∞	13.8	2.3
	good	7.6	∞	10.3	∞	12.3	4.0
	maids	6.6	∞	∞	∞	4.5	3.2
	sleep	6.8	12.6	11.6	∞	10.3	2.0

		the	them	good	maids	sleep
Channel model	les	2.3	1.1	∞	∞	∞
	le	2.1	∞	∞	∞	∞
	la	1.7	∞	∞	∞	∞
	eux	∞	1.4	∞	∞	∞
	elles	∞	2.6	∞	∞	∞
	bonnes	∞	∞	4.0	5.8	∞
	bonne	∞	∞	1.8	∞	∞
	honnête	∞	∞	6.6	∞	∞
	servantes	∞	∞	∞	3.3	∞
	dormir	∞	∞	∞	∞	1.8
	dormez	∞	∞	∞	∞	7.0
	dorment	∞	∞	∞	∞	2.9

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 'bonnes' in state* **maids** *is* `5.8`*. In the language model, full stop ('.') is used for sentence start and dollar-sign ('$') for sentence end.*

The language model costs 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.