UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Thursday 24th May 2012

09:30 to 11:30

Year 3 Courses

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

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 **gold standard** (in the context of speech and language processing)? What are two important uses for a **gold standard** in developing a language-processing system?

2. What is **backoff** and why is it necessary?

3. What is meant by the **edit distance** between two strings? Explain *briefly* how **dynamic programming** can be used to compute it.

4. Assuming that you have developed a POS tagger as a Hidden Markov Model (HMM), write the formula for estimating the probability that the sentence *the man sleeps* is tagged $D\ N\ V$, assuming a tagset T.
   Assume that the transition model provides probabilities for $P(t|s)$ and $P(e|t)$ for each $t \in T$, where s marks the start of a sentence and e its end.

5. What is the main difference between **data-intensive** and **rule-based** approaches to speech and language processing, and what are some of the reasons for the rise of the former approach?

6. What is a **significance test** and why is it important? What is the difference between **parametric** and **non-parametric** tests?

7. What is the formula for **mutual information**? Give an example of a language processing task where it might be used.

8. Identify some of the drawbacks of a **dictionary-style model** of lexical meaning.

Part B

Answer two questions in Part B.

1. **HMM POS Tagging**

An HMM POS tagger consists of a *transition model* (also sometimes known as a *language model*) and a *sensor model* (also sometimes known as the *channel model*) shown in the tables below. These show the *cost*—that is, negative log probabilities—so that (a) you can sum them rather than multiplying, and (b) you are looking to minimise the total cost.

Transition model: Gives the conditional cost distribution $-log(P(t_i|t_{i-1}))$, with t_{i-1} down the left and t_i across the top. '.' and \$ are used respectively for sentence start and sentence end.

$t_{i-1}\backslash t_i$	D	N	V	P	A	\$
.	0.7	3.3	4.3	5	2.3	∞
D	∞	1.3	∞	5	1.7	∞
N	4	2.3	0.7	1.3	7	0.3
V	0.3	1.3	∞	2.3	3.3	0.7
P	0.3	3.3	4	5	3.3	5
A	∞	1.3	5	∞	2.3	4

Sensor model: Gives the conditional cost distribution $-log(P(w_i|t_i))$.

	D	N	V	P	A
the	0.0	∞	∞	∞	∞
dog	∞	0.2	4	∞	∞
jumped	∞	∞	0.2	∞	5
on	∞	∞	∞	0.0	∞
cotton	∞	0.2	4	∞	1.2
rug	∞	0.0	∞	∞	∞

(a) Derive the equation for estimating the most likely tag sequence t_1^n of a word sequence w_1^n according to the HMM model for POS tagging. Give your answer both in terms of probabilities and in terms of costs (negative log probabilities). [*2 marks*]

(b) Describe in detail the assumptions about conditional independence that are made in the answer you gave in part (a), and for each independence assumption demonstrate by example that while the assumption may be practical, it is not always valid. [*3 marks*]

(c) Consider the following sentence:

The dog jumped on the cotton rug.

i. Give the likelihood, in terms of cost, that this sentence receives the POS tag sequence DNVPDAN, using the costs in tables above. [*5 marks*]

ii. Is tagging *cotton* as an A or an N more likely, assuming that the tags for the other words in the sentence remain as in (c)i above? Justify your answer. [*3 marks*]

2. **Chart parsing**

(a) Describe in detail bottom-up depth-first chart parsing using context-free phrase structure grammars.

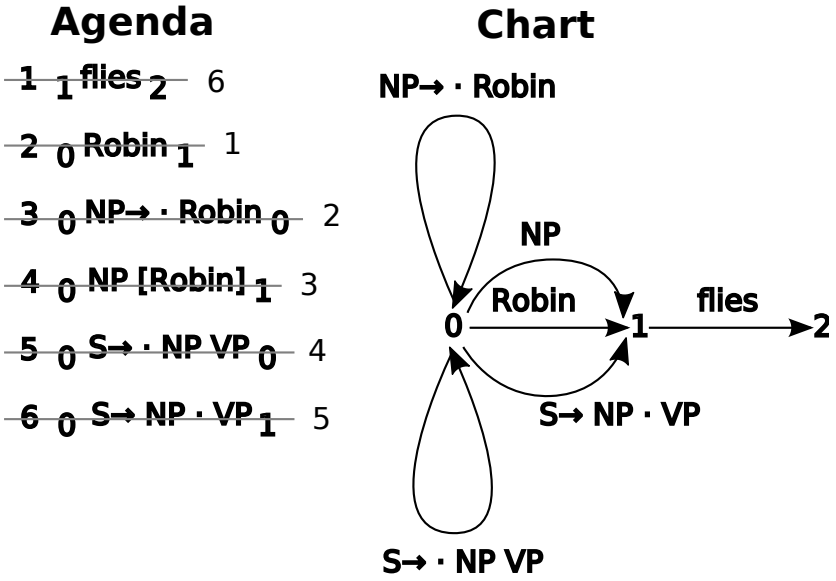
Include descriptions of the chart and its constituent parts, the agenda and the grammar and the four different processing rules which create edges.

Be sure to describe the overall process of parsing: how it starts, how the agenda and chart interact, how it finishes. [*6 marks*]

(b) Here is a simple context-free grammar:

R1: S ⟶ NP VP
R2: VP ⟶ V0
R3: VP ⟶ V1 NP
R4: V0 ⟶ *snores* | *flies* | ...
R5: V1 ⟶ *likes* | *flies* | ...
R6: NP ⟶ *Robin* | ...

And a snapshot of the agenda and chart part-way through a bottom-up depth-first parse of "Robin flies" using that grammar. The agenda entries are numbered to show the order the edges were added. The *strike-throughs* are numbered to show the order in which the edges came *off* the agenda and into the chart.



For the six edges that were on the agenda, now in the chart, identify the processing rule which created them and, in each case, the inputs involved, e.g. `FR(n,m)` for the Fundamental Rule operating on edges `m` and `n` [*3 marks*]

(c) Complete the parse. Copy the above diagram of the chart (you don't need to copy the agenda), and show *all* the subsequent changes to the agenda and the chart. Only one drawing is necessary: use numbered strike-throughs as above to show the order in which edges come off the agenda and into the chart. For example

~~8 FR(3,1) 0 D [the] 1~~ **3**

would mean that edge 8 was the third to be taken off the agenda into the chart. Continue to number the edges on the agenda, starting with `7`. You should also identify the processing rule and inputs involved for each edge on the agenda, as above. [*4 marks*]

3. **Lexical Semantics**

(a) Describe the Lesk algorithm for word sense disambiguation. [*4 marks*]

(b) Consider the following WordNet definitions:

WordNet definitions: For the senses of the noun *dock*.

sn1: dock (an enclosure in a court of law where the defendant sits during the trial).

sn2: dock, sorrel, sour grass (any of certain coarse weedy plants with long taproots, sometimes used as table greens or in folk medicine).

sn3: pier, wharf, wharfage, dock (a platform built out from the shore into the water and supported by piles; provides access to ships and boats).

sn4: dock, loading dock (a platform where trucks or trains can be loaded or unloaded).

sn5: dock, dockage, docking facility (landing in a harbour next to a pier where ships are loaded and unloaded or repaired; may have gates to let water in or out) *the ship arrived at the dock more than a day late.*

sn6: dock (the solid bony part of the tail of an animal as distinguished from the hair).

sn7: bobtail, bob, dock (a short or shortened tail of certain animals).

Assuming the WordNet definitions for the noun *dock* given above, and assuming that a neighbourhood for a word consists of all words in the same sentence as the target word, its preceding three sentences and its next three sentences in the text, simulate the Lesk algorithm you gave in answer to part (a) to predict the sense assigned to *dock* in the following text.
Hint: use the **lemmas** *of the open-class words in the definitions and the text rather than the words themselves.*

> The judge at the trial asked the barrister where the defendant was.
> The barrister provided the information that there was a rumour going round that the criminal was attempting to escape by ship from the harbour at Dover.
> The barrister then went to Dover to look for him.
> But the defendant was eventually found slumped unconscious at the **dock**.

Does this match the correct sense for *dock* in this example? Explain your answer. [*4 marks*]

(c) Some of the senses given above are related. So let's attempt to make word sense disambiguation a bit easier by collapsing each set of related senses into a single sense, and defining that single sense by simply conjoining the words used to describe the related senses. Identify how the senses listed above would be clustered into new (and fewer) senses. Justify your answer by explaining how the senses in each new group are related. [*3 marks*]

(d) With your new sense definitions for *dock* given in your answer to part (c), use the Lesk algorithm to predict which sense (or senses) would be assigned to *dock* in the above text. Is this sense the correct one? [*2 marks*]