UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR09028 FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Friday 15th May 2015

14:30 to 16:30

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 part A and one book for each question in Part B; that is, three books in all.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

Year 3 Courses

Convener: S. Viglas
External Examiners: A. Cohn, T. Field

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

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 one or two 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**? Draw a graph showing the most common frequency distribution of natural language items. What is its name? Give two examples of items that have this distribution.

2. What is a **gold standard** (in the context of speech and language processing)? Why is **gold standard** data often divided into two parts when developing a language processing system?

3. Name and define two types of **syntactic ambiguity**, give a linguistic example of each, and briefly explain why they present a major challenge for **parsing**.

4. Name two types of linguistic ambiguity that a POS tagger cannot completely eliminate, even in principle, and explain why not.

5. Assuming a **trigram 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 saw Kim dance" made up of simpler probabilities.

6. Assuming a **POS tagger** as a Hidden Markov Model (HMM), give the probability that the phrase "I saw Kim dance" is tagged PRO VBD NNP VB in terms of probabilities from the transition model and sensor model of the HMM.
   Note: assume that the start and end of a sentence is marked $\langle s \rangle$ and $\langle /s \rangle$ respectively.

7. What problem are **backoff** and **smoothing** trying to solve? Briefly describe how each of them works.

8. Assume that you are building a **word sense disambiguation** model that represents the context of each occurrence of the target word as a vector, consisting of the lemmas of the previous two open class words and the lemmas of the next open class word, in that order. Then if the target word is *market*, what would the vector be for representing its context in the sentence "The stocks went up as much as the market expected them in the current climate"?

$t_{i-1} \backslash t_i$	NN	VB	JJ	DT	$\langle /s \rangle$
NN	20.6	3.1	45.0	10.0	3.0
VB	15.0	20.0	20.0	2.0	2.0
JJ	9.0	35.0	8.0	45.0	25.0
DT	6.0	45.0	3.0	50.0	55.0
$\langle s \rangle$	8.0	30.0	20.0	2.0	∞

Table 1: Transition Model

$t \backslash w$	*the*	*boy*	*likes*	*boring*	*relatives*
NN	70.0	2.0	4.0	15.0	2.0
VB	∞	15.0	2.0	2.0	25.0
JJ	∞	∞	45.0	8.0	30.0
DT	2.0	∞	∞	∞	∞

Table 2: Sensor Model

Part B

Answer two questions in Part B.

1. **Dynamic Programming and Hidden Markov Models**

   (a) What independence assumptions are made by a Hidden Markov Model POS tagger? [*1 mark*]

   (b) Assuming a set T of POS tags, a tag $\langle s \rangle$ for the start of a sentence and a tag $\langle /s \rangle$ for the end of a sentence, derive the Hidden Markov Model (HMM) formula for identifying the most likely tag sequence t_1^n of a word sequence w_1^n. Derive the formula using probabilities, and justify each step in the derivation. Then state the derived formula using *costs*—that is, negative log probabilities. [*4 marks*]

   (c) Assuming that the POS tagger is defined by the transition and sensor models in Tables 1 and 2, give the negative log probability C that the sentence "the boy likes boring relatives" receives the POS tag sequence DT, NN, VB, JJ, NN.
   Note: All numbers are costs—that is, negative log probabilities. So you can sum them rather than multiply them. [*5 marks*]

   (d) What happens when tagging "teh boy likes boring relatives" with the channel and sensor models in Tables 1 and 2? Describe briefly how one might meet the challenge of POS tagging unrestricted and unedited text. [*3 marks*]

R1: S ⟶ NP VP
R2: N ⟶ *ladies* | *dance* | ...
R3: MOD ⟶ *ladies* | ...
R4: DET ⟶ *the* | ...
R5: N ⟶ MOD N
R6: NP ⟶ DET N
R7: VP ⟶ V0
R8: V0 ⟶ *dance* | ...

Figure 1: A simple context free grammar

2. **Chart Parsing**

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

Include descriptions of the chart and its constituent parts, the agenda, the grammar, the input and the processing rules which create edges: that is, the lexical rule, the bottom up rule and the fundamental rule.

Be sure to describe the overall process of parsing: how it starts, how the agenda and chart interact, how it finishes. [*5 marks*]

(b) Figure 1 shows a simple context-free grammar. Figure 2 shows the edges that were put onto the agenda (in the order they were put onto it) part-way through the bottom up depth-first parse of "the ladies dance" using that grammar. Figure 3 shows the chart that exists part-way through the bottom-up parse of "the ladies dance".

Note: We have not shown the order in which the edges *came off* the agenda and were put onto the chart! Rather, the entries are numbered to show the order in which the edges were added to the agenda. The rules which created the edge are also shown: e.g., `FR(n,m)` means that the current edge was added to the agenda as a result of the Fundamental Rule operating on edges `m` and `n`; similarly `BU(m,n)` stands for the Bottom Up rule and `L` for the lexical rule.

Complete the parse. Write down *all* the subsequent edges that are added to the agenda (you don't need to copy down the edges 1–12 above), in the order in which they are added. Show this by continuing to number the edges on the agenda, starting with `13`. You should also identify for each edge the processing rule and inputs that create it, as illustrated above (e.g., `FR(10,11)`). Then copy the above chart, and complete it by showing how the edges that you have added to the agenda get added to the chart. You need only do one drawing, showing the final chart.

You don't have to depict active arcs as dotted lines and inactive ones as dashed lines (as done in Figure 3), but be sure to label each arc clearly! [*8 marks*]

1L:	dance [2,3]
2L:	ladies [1,2]
3L:	the [0,1]
4 BU(3,R4):	DET $\longrightarrow$ • the [0,0]
5 FR(3,4):	DET $\longrightarrow$ the • [0,1]
6 BU(5,R6):	NP $\longrightarrow$ • DET N [0,0]
7 FR(5,6):	NP $\longrightarrow$ DET • N [0,1]
8 BU(2,R2):	N $\longrightarrow$ • ladies [1,1]
9 BU(2,R3):	MOD $\longrightarrow$ • ladies [1,1]
10 FR(2,9):	MOD $\longrightarrow$ ladies • [1,2]
11 BU(9,R5):	N $\longrightarrow$ • MOD N [1,1]
12 FR(10,11):	N $\longrightarrow$ MOD • N [1,2]

Figure 2: Edges that have been put onto the agenda part-way through the parse of "the ladies dance", using the grammar from Figure 1.

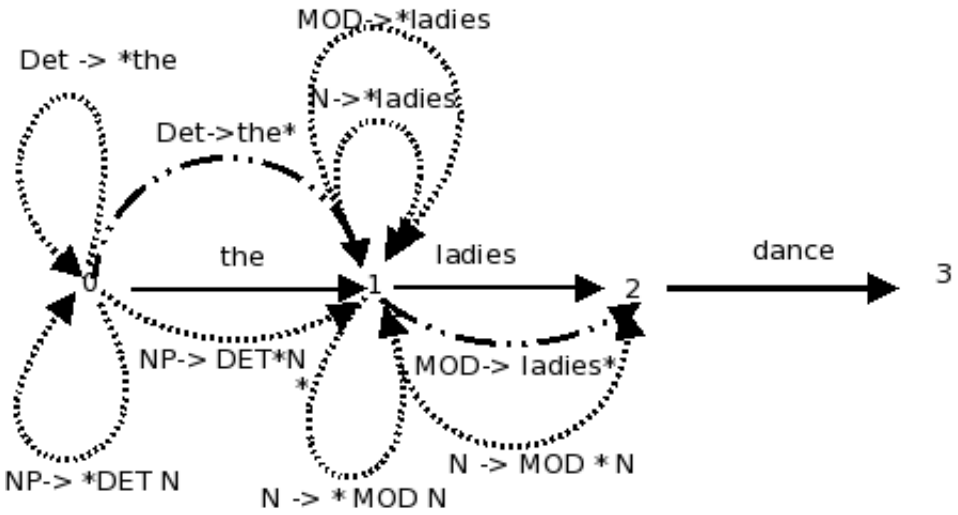

Figure 3: The chart, part-way through parsing "the ladies dance" using the grammar from Figure 1.

3. **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 **magazine** has the following 6 noun senses:

Nouns:

sn1: magazine, mag
(a periodic publication containing pictures and stories and articles of interest to those who purchase it or subscribe to it)
it takes several years before a magazine starts to break even or make money

sn2: magazine
(product consisting of a paperback periodic publication as a physical object)
tripped over a pile of magazines

sn3: magazine, magazine publisher
(a business firm that publishes magazines)
he works for a magazine

sn4: magazine, cartridge
(a light-tight supply chamber holding the film and supplying it for exposure as required)

sn5: magazine, powder store, powder magazine
(a storehouse (as a compartment on a warship) where weapons and ammunition are stored)

sn6: cartridge holder, cartridge clip, clip, magazine
(a metal frame or container holding cartridges; can be inserted into an automatic gun)

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*]