UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR09028 FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Thursday 18 th May 2017

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.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

Year 3 Courses

Convener: C. Stirling
External Examiners: A. Cohn, A. Donaldson, S. Kalvala

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

Part A

Answer ALL questions in Part A.

Your answers should be thorough but they need not be lengthy: from a few words in some cases up to a short paragraph in others. Each question is worth three marks, 24 marks in total for this section.

1. What is a **frequency distribution**? Draw simple sketches of a **normal** distribution and a **Zipf's law** distribution. **Parametric** statistics are appropriate for testing which one for significance?

2. In implementing probabilistic models, we often compute with **negative log probabilities** (costs) instead of actual probabilities. What are the benefits to doing so? What kind of operation is it *not* suitable for?

3. What is the fundamental flaw in a **Probabilistic Context-Free Grammar** constructed using **maximum likelihood estimation** based on a treebank such as the Penn treebank with respect to attachment ambiguity? Illustrate your answer with parse sketches for two simple verb-phrases for the POS tag sequence `V D N Prep D N`.

4. Name two of the three main approaches to **dependency parsing** and give brief descriptions of how they operate.

5. Imagine that you would like to determine how similar two different words are, using **distributional lexical semantics**. Given feature vectors for words w_1 and w_2, what are two different distance or similarity functions you could use, not including **dot product**?

6. Give formulae for the **Precision** and **Recall** evaluation measures using appropriate combinations of **T** and **F** (for "true" and "false") and **P** and **N** (for "positive" and "negative"). Which one will be misleadingly high for a baseline system which always says "Yes" in a binary forced-choice task such as "Is [some stimulus] an English word?" ?

7. If we're interested in knowing how similar the meanings of "the" and "is" are, compared to the meanings of "hyena" and "coyote", why might it be a bad idea to use **dot product** as a similarity function if we're using context word counts for our features?

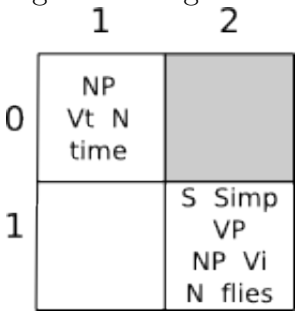
8. What is the main difference between the **PropBank** and **FrameNet** approach to **semantic roles**? Illustrate your answer by contrasting the role labels each of them would assign to the noun *hay* in the sentence *Robin loaded hay onto the cart.*

Part B

Answer two questions in Part B.

1. **Parsing with context-free grammars**

   (a) Cocke-Kasami-Younger (CKY) parsing
   The figure below shows an intermediate state of a CKY parse of the ambiguous string "the frogs fish for fish".



   i. What are the rules that
      - must be in the grammar used in this parse so far
      - must be added to get it to a successful conclusion? [*2 marks*]

   ii. What happens next? That is, what gets inserted into the 6 grey cells in the above figure, in what order? Why? That is, explain in each case how the CKY algorithm determines what does or doesn't get added. [*4 marks*]

   Note that you are *not* required to stick to Chomsky Normal Form for your rules. That is, you may use any symbols on the right-hand side of rules, as long as there are only one or two of them.

   (b) Probabilistic context-free grammar (PCFG)

   i. Give the formula for the **maximum likelihood estimate** of the (inner) probability of a context-free grammar rule $NT \longrightarrow C_1, C_2 \ldots C_n$ given a treebank of parsed sentences. [*1 mark*]

   ii. How does the size of the set of available labels (terminal plus non-terminal symbols) affect the usefulness of a probabilistic grammar built up in this way? [*2 marks*]

   iii. **Lexicalisation** and **dependency grammar** are two possible approaches to solving the problem you identified in Question A3. Pick one, describe its key properties, and explain how it might be used to address the problem. Refer to your examples in your answer. [*4 marks*]

2. **A real noisy channel**

The table below is an extract from the results of a real noisy channel perception experiment, where capital letters where presented for 200 milliseconds and subjects had to report what they saw.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	...	Total
A	168	1	0	2	5	5	1	3	3	1	6	4	0	4	0	2	0	...	242
B	0	136	1	0	3	2	0	4	0	2	4	4	1	0	0	3	1	...	179
C	1	6	111	5	11	6	36	5	5	2	6	11	0	3	7	2	12	...	257
D	1	17	4	157	6	11	0	5	8	2	2	9	0	2	0	30	1	...	280
E	2	10	0	1	98	27	1	5	2	0	2	21	0	0	0	3	0	...	195
F	1	0	0	1	9	73	0	6	5	0	6	4	0	1	0	3	0	...	133
G	1	3	32	1	5	3	127	3	1	0	2	2	1	3	5	2	4	...	223
H	2	0	0	0	3	3	0	44	7	2	0	7	6	7	0	1	0	...	96
I	1	0	0	0	1	3	0	0	52	2	3	5	0	2	0	1	0	...	87
J	1	1	2	3	4	1	0	5	10	85	0	3	7	3	2	0	0	...	157
K	6	3	0	0	3	5	0	7	5	0	90	4	3	3	0	1	0	...	188
L	0	0	0	1	4	1	0	4	10	0	3	63	1	3	0	0	0	...	103
M	0	1	0	0	0	2	0	12	5	1	1	2	113	6	0	0	0	...	161
N	2	1	0	1	6	5	1	42	8	2	6	11	26	117	0	0	0	...	279
O	2	4	39	9	1	3	24	1	3	2	1	4	0	0	174	3	110	...	403
P	0	4	1	3	3	11	1	3	6	1	4	3	0	1	0	117	0	...	202
Q	0	0	9	2	0	1	4	1	0	2	0	0	0	0	8	0	71	...	106
R	2	3	0	0	8	9	2	3	6	1	6	4	1	1	0	19	0	...	213
S	4	6	0	1	6	0	2	0	2	3	5	2	0	2	1	4	0	...	212
T	0	0	0	0	7	13	0	7	30	2	10	12	1	2	0	4	0	...	230
U	0	2	1	11	3	4	1	31	16	76	6	14	12	17	1	1	0	...	374
V	2	2	0	1	1	1	0	0	6	8	9	2	10	4	0	1	0	...	237
W	0	0	0	1	1	4	0	4	5	2	7	2	15	10	2	1	0	...	232
X	2	0	0	0	2	3	0	3	0	2	9	2	0	4	0	0	1	...	93
Y	0	0	0	0	1	1	0	2	3	0	8	0	2	1	0	0	0	...	124
Z	2	0	0	0	9	3	0	0	2	2	4	5	1	4	0	2	0	...	194
Total	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200	200		

(a) What is such a table called? What does the value `110` at the intersection of the **Q** column and the **O** row signify? What is the **maximum likelihood estimate** of the probability of someone reporting a **D** when shown a **P** in the same experimental conditions? [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

(b) The following table transforms the upper-left-hand corner of the table on the previous page into negative log (base 2) probabilities, rounded to the nearest integer:

	A	B	C	D	E	F
A	0	8	∞	7	5	5
B	∞	1	8	∞	6	7
C	8	5	1	5	4	5
D	8	4	6	0	5	4
E	7	4	∞	8	1	3
F	8	∞	∞	8	4	1

Using this information alone, if a subject reports seeing **CCC**, which of *BED*, *DAB* or *BEE* is the most likely original stimulus? Show your work. [*3 marks*]

(c) Suppose now you have the additional information given below for the cost of the relevant bigrams in English.

$c_{i-1} \backslash c_i$	A	B	C	D	E	F
A	13	6	5	5	10	7
B	4	7	11	11	2	16
C	3	13	6	12	3	13
D	5	11	11	7	3	11
E	4	9	5	4	5	7
F	4	12	15	14	4	4

Given a report of **CBD** this time, draw a 6-by-3 matrix with rows labelled **A**–**F** from top to bottom and column labels **C B D** and fill it in with an implementation of **Viterbi search** to use this new information to find the most likely English 3-letter-word input. [*5 marks*]

(d) What's missing from the bigram table and the Viterbi matrix which would probably give better results? [*2 marks*]

3. **Lexical Semantics**

(a) Explain the difference between homonymy and polysemy. Illustrate your answer with examples of each kind of ambiguity. [*2 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, argue, 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*

Identify the pair of noun senses that are most similar to one another, and the pair of verb senses that are least similar, and briefly explain your choices. [*3 marks*]

(c) Suppose you have annotated 5 percent of the Penn treebank corpus with WordNet senses as above, and want to disambiguate the different senses of the word *scrap* in a syntactically annotated sample text (which has no sense labels yet) using a maximum entropy model. What type of feature could you use? Give a specific example feature of that type. [*2 marks*]

(d) Give an example of a different feature that would become more useful if you had 1000 times as many annotated sentences. Explain why it would become more useful. [*2 marks*]

QUESTION CONTINUES ON NEXT PAGE

(e) Suppose that we have trained our classifier using features $f_{1\ldots n}$. We now want to know the proportional probabilities of two senses: $\frac{P(S=sv1|\vec{x})}{P(S=sv2|\vec{x})}$, where S is the sense of the instance of *scrap* that we're currently considering. Suppose we only have features based on the template `word-contains-letter(`l`) &` s, where s is the sense and l is a case-insensitive Latin letter. Derive an expression for how the ratio changes when we remove all instances of the letter 'a' from the word. Show your work. [*4 marks*]