THIS PAPER IS NOT TO BE REMOVED FROM THE EXAMINATION HALLS

UNIVERSITY OF LONDON **CO3354 ZA**

BSc Examination

COMPUTING AND INFORMATION SYSTEMS, CREATIVE COMPUTING and COMBINED DEGREE SCHEME

Introduction to Natural Language Processing

Date and Time: Monday 15 May 2017: 14:30 - 16:45

Duration: 2 hours 15 minutes

There are **FIVE** questions on this paper. Candidates should answer **THREE** questions. All questions carry equal marks and full marks can be obtained for complete answers to **THREE** questions. The marks for each part of a question are indicated at the end of the part in [.] brackets.

Only your first **THREE** answers, in the order that they appear in your answer book, will be marked.

There are 75 marks available on this paper.

A handheld calculator may be used when answering questions on this paper but it must not be pre-programmed or able to display graphics, text or algebraic equations. The make and type of machine must be stated clearly on the front cover of the answer book.

QUESTION 1

Consider the following grammar:

S → NP VP
NP → Det N
NP → PN
NP → NP PP
VP → V
VP → V NP
VP → V NP PP
PP → P NP

Det → the | a
N → train | carriage | seat | window | station | pipe | newspaper
PN → Holmes | Watson | Moriarty | Lestrade
V → slept | sat | reserved | smoked | read | saw | met
P → in | on | under | by | at

a)

i. Write out two grammatical sentences of at least 10 words and with different structures, which are generated by this set of rules. **[2]**

ii. Draw syntax trees for both of these sentences, according to the grammar rules above. You should draw all applicable trees if your sentences are structurally ambiguous. **[6]**

b) Explain how the grammar can be modified so that it will generate the grammatical examples (i-v) below. You should:

- Identify the constructions which are not covered in the original grammar. **[4]**
- Propose new or modified rules to handle these constructions, with appropriate worked examples. **[5]**

i. Holmes sat by the window and slept.
ii. Holmes and Watson reserved a compartment on the train.
iii. Holmes read the newspaper and a telegram.
iv. Moriarty wrote a long and difficult monograph.
v. Moriarty wrote a difficult long monograph.

c)

i. What problem could this grammar (prior to your modifications) cause for a recursive-descent parser? **[3]**

ii. Explain how the rules could be modified to get round this problem. **[3]**

iii. What effect would your modification have on the coverage of the grammar? **[2]**

QUESTION 2

A probabilistic phrase structure grammar

```
S     -> NP VP                [0.8]
S     -> NP Adv VP            [0.2]
VP    -> Adv VP               [0.2]
VP    -> VP 'or' VP           [0.1]
VP    -> V                    [0.7]
NP    -> 'Maria'              [1.0]
V     -> 'runs'               [0.5]
V     -> 'swims'              [0.5]
Adv   -> 'regularly'          [1.0]
```

a) The rules shown above make up an example of a probabilistic or weighted grammar. What advantages can such grammars have over conventional phrase-structure grammars?

[6]

b) Using the probabilistic grammar rules and lexical rules given above, draw as many syntax trees as you can for the sentence:

"Maria regularly runs or swims".

Explain any ambiguities it may have by giving paraphrases for the meanings corresponding to the different syntactic analyses.

[9]

c)

i. Calculate the relative probabilities assigned to different analyses of the sentence by the grammar rules. Which analysis has the highest probability? **[6]**

ii. Discuss whether the results agree with your intuitive understanding of the sentence.

[4]

QUESTION 3

The following paragraph is taken from the first chapter of Grant Allen's *An African Millionaire* (1897). (Note that a 'sharper' in this context means 'a swindler, a professional gambler'.)

My name is Seymour Wilbraham Wentworth. I am brother-in-law and secretary to Sir Charles Vandrift, the South African millionaire and famous financier. Many years ago, when Charlie Vandrift was a small lawyer in Cape Town, I had the (qualified) good fortune to marry his sister. Much later, when the Vandrift estate and farm near Kimberley developed by degrees into the Cloetedorp Golcondas, Limited, my brother-in-law offered me the not unremunerative post of secretary; in which capacity I have ever since been his constant and attached companion.

He is not a man whom any common sharper can take in, is Charles Vandrift. Middle height, square build, firm mouth, keen eyes--the very picture of a sharp and successful business genius.

This is the result of running the above text through the Lancaster Stemmer:

```
['my', 'nam', 'is', 'seymo', 'wilbraham', 'wentwor', '.', 'i', 'am',
'brother-in-law', 'and', 'secret', 'to', 'sir', 'charl', 'vandrift',
',', 'the', 'sou', 'afr', 'millionair', 'and', 'fam', 'fin', '.',
'many', 'year', 'ago', ',', 'when', 'char', 'vandrift', 'was', 'a',
'smal', 'lawy', 'in', 'cap', 'town', ',', 'i', 'had', 'the', '(',
'qual', ')', 'good', 'fortun', 'to', 'marry', 'his', 'sist', '.',
'much', 'lat', ',', 'when', 'the', 'vandrift', 'est', 'and', 'farm',
'near', 'kimberley', 'develop', 'by', 'degr', 'into', 'the',
'cloetedorp', 'golconda', ',', 'limit', ',', 'my', 'brother-in-law',
'off', 'me', 'the', 'not', 'unremun', 'post', 'of', 'secret', ';',
'in', 'which', 'capac', 'i', 'hav', 'ev', 'sint', 'been', 'his',
'const', 'and', 'attach', 'comp', '.', 'he', 'is', 'not', 'a', 'man',
'whom', 'any', 'common', 'sharp', 'can', 'tak', 'in', ',', 'is',
'charl', 'vandrift', '.', 'middl', 'height', ',', 'squ', 'build',
',', 'firm', 'mou', ',', 'keen', 'ey', '--', 'the', 'very', 'pict',
'of', 'a', 'sharp', 'and', 'success', 'busy', 'geni', '.']
```

And this is the result of running it through the Porter Stemmer:

```
['My', 'name', 'is', 'Seymour', 'Wilbraham', 'Wentworth', '.', 'I',
'am', 'brother-in-law', 'and', 'secretari', 'to', 'Sir', 'Charl',
'Vandrift', ',', 'the', 'South', 'African', 'millionair', 'and',
'famou', 'financi', '.', 'Mani', 'year', 'ago', ',', 'when',
'Charli', 'Vandrift', 'wa', 'a', 'small', 'lawyer', 'in', 'Cape',
'Town', ',', 'I', 'had', 'the', '(', 'qualifi', ')', 'good',
'fortun', 'to', 'marri', 'hi', 'sister', '.', 'Much', 'later', ',',
'when', 'the', 'Vandrift', 'estat', 'and', 'farm', 'near',
'Kimberley', 'develop', 'by', 'degre', 'into', 'the', 'Cloetedorp',
'Golconda', ',', 'Limit', ',', 'my', 'brother-in-law', 'offer', 'me',
'the', 'not', 'unremun', 'post', 'of', 'secretari', ';', 'in',
'which', 'capac', 'I', 'have', 'ever', 'sinc', 'been', 'hi',
'constant', 'and', 'attach', 'companion', '.', 'He', 'is', 'not',
'a', 'man', 'whom', 'ani', 'common', 'sharper', 'can', 'take', 'in',
',', 'is', 'Charl', 'Vandrift', '.', 'Middl', 'height', ',', 'squar',
'build', ',', 'firm', 'mouth', ',', 'keen', 'eye', '--', 'the',
'veri', 'pictur', 'of', 'a', 'sharp', 'and', 'success', 'busi',
'geniu', '.']
```

a)

i. Explain what is meant by **word stems**, with reference to examples from the above text.

ii. Give an example of how stemming can be useful for information retrieval applications.

iii. Explain the difference between a **stemmer** and a **lemmatizer**.

[9]

b)

i. Make a list of rules which the Lancaster stemmer seems to have applied in this example and discuss the motivations for the rules.

[6]

ii. Make a similar list for the Porter stemmer and note any cases where the two stemmers have different results.

[6]

c) Are there any cases where you think the rules were applied incorrectly? If so, give up to four examples. Justify your answer.

[4]

QUESTION 4

a) The following sentences are all ambiguous in some way. Express their different meanings using paraphrases and explain the source of the ambiguity in each case:

i. The professor said there would be an exam on Friday.
ii. I saw her duck.
iii. Ludwig asked Bertrand if he was a complete idiot.
iv. Cecil likes to read romantic novels and cookbooks.

[4 x 1.5]

b) Explain what is meant by a **corpus** in the context of Natural Language Processing.

[2]

c) Briefly describe the following linguistic resources (no more than two sentences each).

i. Project Gutenberg
ii. Penn Treebank
iii. Wordnet
iv. The Bank of English

[4 x 2]

d) Annotate the text below with POS (part of speech) tags, using the universal tagset given in Appendix A as in this example:

```
[('A', 'DET'), ('study', 'NOUN'), ('led', 'VERB'), ('by',
'ADP'), ('Dr', 'NOUN'), ('Kelly', 'NOUN'), ('Jakubowski',
'NOUN'), ('discovered', 'VERB'), ('that', 'ADP'),
('‘earworms’', 'NOUN'), ('are', 'VERB'), ('usually', 'ADV'),
('faster', 'ADJ'), (',', '.'), ('with', 'ADP'), ('a', 'DET'),
('fairly', 'ADV'), ('generic', 'ADJ'), ('melody', 'NOUN')]
```

(earworm = catchy pop song or melody)

Where a word has more than one possible POS, explain how you have decided which one to use.

Text (Goldsmiths Website):

Three of the top 10 most frequently named catchy tracks in the study were by Lady Gaga.

Dr Jakubowski, who completed her PhD research at Goldsmiths and has since joined Durham University’s Department of Music, also confirmed the assumption that songs that get more radio time are more likely to be reported as involuntary musical imagery, or earworms.

[9]

QUESTION 5

a) Briefly explain what is meant by each of the following in the context of Natural Language Processing:

i. Constituent structure
ii. Tokenisation
iii. Opinion mining
iv. N-gram tagging

[4 x 2]

b) Using the regular expressions supplied in Appendix B, describe and give examples to illustrate the classes of strings matched by the following regular expressions (for example, (ac)* matches ε, ac, acac ...):

i.	(ab|c)*	**[1.5]**
ii.	a+|b	**[1.5]**
iii.	[A-Z]?[0-9]+	**[3]**
iv.	([aeiou]+|[0-9]*)	**[3]**

c) Suppose a corpus contains 2m word-tokens, and 400,000 of these are tagged as V (verb). The word-form *walk* occurs 2,000 times in the corpus, tagged either as N or V. Analysis shows that *walk* accounts for 0.4% of all verb tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of *walk* is tagged as V. Explain your answer and show your working. Show your final and intermediate results to no more than two significant figures.

[8]

APPENDIX A: UNIVERSAL PART-OF-SPEECH TAGSET

Tag	Meaning	English Examples
ADJ	adjective	new, good, high, special, big, local
ADP	adposition	on, of, at, with, by, into, under
ADV	adverb	really, already, still, early, now
CONJ	conjunction	and, or, but, if, while, although
DET	determiner, article	the, a, some, most, every, no, which
NOUN	noun	year, home, costs, time, Africa
NUM	numeral	twenty-four, fourth, 1991, 14:24
PRT	particle	at, on, out, over per, that, up, with
PRON	pronoun	he, their, her, its, my, I, us
VERB	verb	is, say, told, given, playing, would
.	punctuation marks	. , ; !
X	other	*ersatz, esprit, dunno, gr8, univeristy*

APPENDIX B: REGULAR EXPRESSIONS

.	Wildcard, matches any character
^abc	Matches some pattern abc at the start of a string
abc$	Matches some pattern abc at the end of a string
[abc]	Matches one of a set of characters
[A-Z0-9]	Matches one of a range of characters
[^abc]	Inside brackets [.], caret is a negation operator
ed|ing|s	Matches one of the specified strings (disjunction)
*	Zero or more of previous item, e.g. a*, [a-z]*
+	One or more of previous item, e.g. a+, [a-z]+
?	Zero or one of the previous item (i.e. optional), e.g. a?, [a-z]?
a(b|c)+	Parentheses that indicate the scope of the operators

END OF PAPER