THIS PAPER IS NOT TO BE REMOVED FROM THE EXAMINATION HALLS

UNIVERSITY OF LONDON **CO3354 ZB**

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 → police | constable | handcuffs | burglar | bed | telephone | gun | table | garden | knife
PN → Holmes | Watson | Moriarty | Lestrade
V → slept | put | saw | called | gave | has | have | arrested
P → in | on | under | by

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) This grammar will generate sequences that are not grammatical sentences. Explain how it can be modified so that it will generate the grammatical examples (i-iii) below but not the ungrammatical (iv-vi). 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. The constable arrested a burglar.
ii. Watson put the knife on the table.
iii. Holmes saw the knife on the table.
iv. *The constable arrested.
v. *Watson put the knife.
vi. *Holmes saw on the table.

c)

i. What problem could this grammar 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) 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. (a|b)* **[1.5]**
ii. a|b+ **[1.5]**
iii. [a-z]+[0-9]+ **[3]**
iv. ([^aeiou]|[0-9])* **[3]**

c) Suppose a corpus contains 500,000 word-tokens, and 125,000 of these are tagged as N (common noun). The word-form *attempt* occurs 10,000 times in the corpus, tagged either as N or V. Analysis shows that *attempt* accounts for 0.3% of all common noun tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of *attempt* is tagged as N. Explain your answer and show your working. Show your final and intermediate results to no more than two significant figures.

[8]

QUESTION 3

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. Use one onion and two fresh chilis or half a teaspoon of chili powder.
   ii. Every student did not pass the exam.
   iii. Fred drove his motorbike into a tree and badly damaged it.
   iv. Smoking seriously harms your health.

[4 x 1.5]

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

[2]

What is meant by the following terms in the context of corpus linguistics? (One mark each for i-viii.)

   i. Monolingual corpus
   ii. Concordancing
   iii. Collocation
   iv. Temporal corpus structure
   v. Development and test sets
   vi. Gold standard
   vii. Inter-annotator agreement
   viii. Stopwords

[8]

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

```
[('With', 'ADP'), ('work', 'NOUN'), ('now', 'ADV'),
('underway', 'ADV'), ('on', 'ADP'), ('Goldsmiths', 'NOUN'),
(',', '.'), ('University', 'NOUN'), ('of', 'ADP'), ('London's',
'NOUN'), ('new', 'ADJ'), ('gallery', 'NOUN'), (',', '.'),
('the', 'DET'), ('hunt', 'NOUN'), ('is', 'VERB'), ('on',
'ADP'), ('for', 'ADP'), ('a', 'DET'), ('Director', 'NOUN'),
('.', '.')]
```

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

Text (Goldsmiths Website):

Goldsmiths is looking for a Director who will build a distinctive and internationally respected programme of exhibitions, residencies and events.

Based within the world-renowned Department of Art, the Director will have responsibility for determining the gallery's programme, in collaboration with the venue's advisory board and steering group. The Director will also hold operational responsibility for fundraising, supported by Department of Art and Development Team.

[9]

QUESTION 4

A probabilistic phrase structure grammar

```
S     -> NP VP              [0.9]
S     -> NP Adv VP          [0.1]
VP    -> Adv VP             [0.2]
VP    -> VP 'and' VP        [0.1]
VP    -> IV                 [0.7]
NP    -> 'Jose'             [1.0]
IV    -> 'sleeps'           [0.5]
IV    -> 'dreams'           [0.5]
Adv   -> 'often'            [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:

"Jose often sleeps and dreams".

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 5

The following paragraph is taken from the first chapter of Edward Gibbon's *Decline and Fall of the Roman Empire* (1776). (Note that 'Aera' is an archaic spelling of 'Era'.)

In the second century of the Christian Aera, the empire of Rome comprehended the fairest part of the earth, and the most civilized portion of mankind. The frontiers of that extensive monarchy were guarded by ancient renown and disciplined valor. The gentle but powerful influence of laws and manners had gradually cemented the union of the provinces. Their peaceful inhabitants enjoyed and abused the advantages of wealth and luxury. The image of a free constitution was preserved with decent reverence: the Roman senate appeared to possess the sovereign authority, and devolved on the emperors all the executive powers of government. During a happy period of more than fourscore years, the public administration was conducted by the virtue and abilities of Nerva, Trajan, Hadrian, and the two Antonines.

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

```
['in', 'the', 'second', 'century', 'of', 'the', 'christian', 'aer',
',', 'the', 'empir', 'of', 'rom', 'comprehend', 'the', 'fairest',
'part', 'of', 'the', 'ear', ',', 'and', 'the', 'most', 'civil',
'port', 'of', 'mankind', '.', 'the', 'fronty', 'of', 'that',
'extend', 'monarchy', 'wer', 'guard', 'by', 'ant', 'renown', 'and',
'disciplin', 'val', '.', 'the', 'gentl', 'but', 'pow', 'influ', 'of',
'law', 'and', 'man', 'had', 'grad', 'cem', 'the', 'un', 'of', 'the',
'provint', '.', 'their', 'peac', 'inhabit', 'enjoy', 'and', 'abus',
'the', 'adv', 'of', 'weal', 'and', 'luxury', '.', 'the', 'im', 'of',
'a', 'fre', 'constitut', 'was', 'preserv', 'with', 'dec', 'rev', ':',
'the', 'rom', 'sen', 'appear', 'to', 'possess', 'the', 'sovereign',
'auth', ',', 'and', 'devolv', 'on', 'the', 'emp', 'al', 'the',
'execut', 'pow', 'of', 'govern', '.', 'dur', 'a', 'happy', 'period',
'of', 'mor', 'than', 'foursc', 'year', ',', 'the', 'publ', 'admin',
'was', 'conduc', 'by', 'the', 'virtu', 'and', 'abl', 'of', 'nerv',
',', 'tras', ',', 'hadr', ',', 'and', 'the', 'two', 'antonin', '.']
```

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

```
['In', 'the', 'second', 'centuri', 'of', 'the', 'Christian', 'Aera',
',', 'the', 'empir', 'of', 'Rome', 'comprehend', 'the', 'fairest',
'part', 'of', 'the', 'earth', ',', 'and', 'the', 'most', 'civil',
'portion', 'of', 'mankind', '.', 'The', 'frontier', 'of', 'that',
'extens', 'monarchi', 'were', 'guard', 'by', 'ancient', 'renown',
'and', 'disciplin', 'valor', '.', 'The', 'gentl', 'but', 'power',
'influenc', 'of', 'law', 'and', 'manner', 'had', 'gradual', 'cement',
'the', 'union', 'of', 'the', 'provinc', '.', 'Their', 'peac',
'inhabit', 'enjoy', 'and', 'abus', 'the', 'advantag', 'of', 'wealth',
'and', 'luxuri', '.', 'The', 'imag', 'of', 'a', 'free', 'constitut',
'wa', 'preserv', 'with', 'decent', 'rever', ':', 'the', 'Roman',
'senat', 'appear', 'to', 'possess', 'the', 'sovereign', 'author',
',', 'and', 'devolv', 'on', 'the', 'emperor', 'all', 'the', 'execut',
'power', 'of', 'govern', '.', 'Dure', 'a', 'happi', 'period', 'of',
'more', 'than', 'fourscor', 'year', ',', 'the', 'public',
'administr', 'wa', 'conduct', 'by', 'the', 'virtu', 'and', 'abil',
'of', 'Nerva', ',', 'Trajan', ',', 'Hadrian', ',', 'and', 'the',
'two', 'Antonin', '.']
```

a)

i. Explain what is meant by **word stems**, with reference to examples from the above text. **[2]**

ii. Give an example of how stemming can be useful in real-world NLP applications. **[3]**

iii. Explain the difference between a **stemmer** and a **lemmatizer**. **[4]**

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]

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