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

Tuesday 15 May 2018: 14.30 – 16.45

Time allowed: 2 hours and 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

The Subject Guide for CO3354 lists a number of potential limitations of statistical Natural Language Processing (NLP), concerning:

1. Computational complexity
2. Sparse data
3. Lack of linguistic insight
4. Cost of supervised learning techniques
5. Dependence on human judgments
6. Restriction to English language

Based on your knowledge and understanding of the course materials, recommended readings and your own independent studies, write an essay addressing the following topics:

- What kinds of problems do these limitations pose for NLP? Give concrete examples where possible.
- What techniques have been developed to solve or work round these problems?
- What new technologies are you aware of that may mitigate these problems?

[25]

QUESTION 2

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 | his
N → throne | table | sword | horse | king | knight | queen | chalice
PN → Arthur | Guinevere | Gawain | Lancelot | Bertilak
V → sat | fought | offered | saw
P → in | on | under | by | at | with | to | beside

a)

i. Write out **TWO** sentences of at least 10 words but with different syntactic structures, which are grammatical and meaningful according to your knowledge of English and 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-iv) 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. Arthur sat on the throne and slept.
ii. Gawain and Lancelot fought Bertilak.
iii. Guinevere sang a long and sad song.
iv. Guinevere sang a long sad song.

c)

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

ii. Explain how the rules could be modified to get around this problem.

iii. What effect would your modification have on the coverage of the grammar?

[8]

NB: you are not required to provide complete grammars in answer to (b) and (c), only to explain how the grammars need to be modified, with appropriate worked examples.

QUESTION 3

A probabilistic phrase structure grammar

S -> NP VP	[1.0]
VP -> VP Adv	[0.3]
VP -> VP 'and' VP	[0.1]
VP -> IV	[0.6]
NP -> 'Alicia'	[1.0]
IV -> 'draws'	[0.5]
IV -> 'paints'	[0.5]
Adv -> 'beautifully'	[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? Explain the purpose of the numbers in square brackets [.].

[9]

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

“Alicia draws and paints beautifully.”

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

[6]

c)

i. Calculate the relative probabilities assigned to different analyses of the sentences by the grammar rules. Which analysis has the highest probability?

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

[10]

QUESTION 4

The following paragraphs are taken from the first chapter of Charles Dickens's *Hard Times* (1897):

'NOW, what I want is, Facts. Teach these boys and girls nothing but Facts. Facts alone are wanted in life. Plant nothing else, and root out everything else. You can only form the minds of reasoning animals upon Facts: nothing else will ever be of any service to them. This is the principle on which I bring up my own children, and this is the principle on which I bring up these children. Stick to Facts, sir!'

The scene was a plain, bare, monotonous vault of a school-room, and the speaker's square forefinger emphasized his observations by underscoring every sentence with a line on the schoolmaster's sleeve. The emphasis was helped by the speaker's square wall of a forehead, which had his eyebrows for its base, while his eyes found commodious cellarage in two dark caves, overshadowed by the wall. The emphasis was helped by the speaker's mouth, which was wide, thin, and hard set.

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

```
['‘', 'now', ',', 'what', 'i', 'want', 'is', ',', 'fact', '.',
'teach', 'thes', 'boy', 'and', 'girl', 'noth', 'but', 'fact', '.',
'fact', 'alon', 'ar', 'want', 'in', 'lif', '.', 'plant', 'noth',
'els', ',', 'and', 'root', 'out', 'everyth', 'els', '.', 'you',
'can', 'on', 'form', 'the', 'mind', 'of', 'reason', 'anim', 'upon',
'fact', ':', 'noth', 'els', 'wil', 'ev', 'be', 'of', 'any', 'serv',
'to', 'them', '.', 'thi', 'is', 'the', 'principl', 'on', 'which',
'i', 'bring', 'up', 'my', 'own', 'childr', ',', 'and', 'thi', 'is',
'the', 'principl', 'on', 'which', 'i', 'bring', 'up', 'thes',
'childr', '.', 'stick', 'to', 'fact', ',', 'sir', '!', '’', 'the',
'scen', 'was', 'a', 'plain', ',', 'bar', ',', 'monoton', 'vault',
'of', 'a', 'school-room', ',', 'and', 'the', 'speak', '’', 's',
'squ', 'foref', 'emphas', 'his', 'observ', 'by', 'undersc', 'every',
'sent', 'with', 'a', 'lin', 'on', 'the', 'schoolmaster', '’', 's',
'sleev', '.', 'the', 'emphas', 'was', 'help', 'by', 'the', 'speak',
'’', 's', 'squ', 'wal', 'of', 'a', 'forehead', ',', 'which', 'had',
'his', 'eyebrow', 'for', 'it', 'bas', ',', 'whil', 'his', 'ey',
'found', 'commody', 'cell', 'in', 'two', 'dark', 'cav', ',',
'overshadow', 'by', 'the', 'wal', '.', 'the', 'emphas', 'was',
'help', 'by', 'the', 'speak', '’', 's', 'mou', ',', 'which', 'was',
'wid', ',', 'thin', ',', 'and', 'hard', 'set', '.']
```

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

```
['`', 'now', ',', 'what', 'I', 'want', 'is', ',', 'fact', '.',
'teach', 'these', 'boy', 'and', 'girl', 'noth', 'but', 'fact', '.',
'fact', 'alon', 'are', 'want', 'in', 'life', '.', 'plant', 'noth',
'els', ',', 'and', 'root', 'out', 'everyth', 'els', '.', 'you',
'can', 'onli', 'form', 'the', 'mind', 'of', 'reason', 'anim', 'upon',
'fact', ':', 'noth', 'els', 'will', 'ever', 'be', 'of', 'ani',
'servic', 'to', 'them', '.', 'thi', 'is', 'the', 'principl', 'on',
'which', 'I', 'bring', 'up', 'my', 'own', 'children', ',', 'and',
'thi', 'is', 'the', 'principl', 'on', 'which', 'I', 'bring', 'up',
'these', 'children', '.', 'stick', 'to', 'fact', ',', 'sir', '!',
'’', 'the', 'scene', 'wa', 'a', 'plain', ',', 'bare', ',', 'monoton',
'vault', 'of', 'a', 'school-room', ',', 'and', 'the', 'speaker', '’',
's', 'squar', 'forefing', 'emphas', 'hi', 'observ', 'by',
'underscor', 'everi', 'sentenc', 'with', 'a', 'line', 'on', 'the',
'schoolmast', '’', 's', 'sleev', '.', 'the', 'emphasi', 'wa', 'help',
'by', 'the', 'speaker', '’', 's', 'squar', 'wall', 'of', 'a',
'forehead', ',', 'which', 'had', 'hi', 'eyebrow', 'for', 'it',
'base', ',', 'while', 'hi', 'eye', 'found', 'commodi', 'cellarag',
'in', 'two', 'dark', 'cave', ',', 'overshadow', 'by', 'the', 'wall',
'.', 'the', 'emphasi', 'wa', 'help', 'by', 'the', 'speaker', '’',
's', 'mouth', ',', 'which', 'wa', 'wide', ',', 'thin', ',', 'and',
'hard', 'set', '.']
```

a)

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

[2]

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

[2]

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

[3]

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.

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

[2x7]

c) Are there any cases where you think the results are not genuine word stems? If so, give up to **FOUR** examples. Justify your answer.

[4]

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

[4x2]

b) Using the regular expressions supplied in the Appendix, 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)*
   ii. a+|b
   iii. [A-Z]?[0-9]+
   iv. ([aeiou]+|[0-9]*)

[8]

c) Suppose a novel contains 189,000 word-tokens, and 45,750 of these are tagged as N (common noun). The word-form *work* occurs 92 times in the novel, tagged either as N or V. Analysis shows that *work* accounts for 0.16% of all common noun tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of *work* is tagged as N. Explain your answer and show your working. Show your final and intermediate results to no more than two significant figures.

[9]

APPENDIX: REGULAR EXPRESSIONS

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

END OF PAPER