THIS PAPER IS NOT TO BE REMOVED FROM THE EXAMINATION HALL

UNIVERSITY OF LONDON

CO3354 ZB

BSc EXAMINATION

COMPUTING AND INFORMATION SYSTEMS, CREATIVE COMPUTING and COMBINED DEGREE SCHEME

Introduction to Natural Language Processing

Wednesday 15 May 2019: 14.30 – 16.45

Time allowed: 2 hours and 15 minutes

DO NOT TURN OVER UNTIL TOLD TO BEGIN

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.

Appendices A-D are attached at the end of this examination 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.

UL19/0343

Question 1: Language in Use

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: For example, *Use two onions and one fresh chili or three dried chilis* is an example of conjunction ambiguity; if you use dried chilis, should you use any onions?

i. Ivan shoved Bill and he fell over.
ii. Gregor and Fatima met at the bank.
iii. Miriam watched the cat with one eye.
iv. Flying planes can be dangerous.

[4x2]

b) The Loebner prize is an annual "Turing Test" competition where computer systems are judged on their ability to pass as a human being in unrestricted conversation. The contest for 2018 was run under the supervision of the AISB (Society for the Study of Artificial Intelligence and Simulation of Behaviour). The contest included a preliminary round where each entrant was presented with a fixed set of 20 questions, and the four highest scoring entrants went on to compete in the final. Appendix A contains two transcripts from the preliminary round: one from the final four, and one which did not score so well. The questions are in *italic* and the system's responses are in normal font.

For this question you should compare the performance of these two entrants. Your answer should address such issues as:

- What can we learn from these examples about the challenges involved in processing natural language communication beyond the level of the sentence?
- What particular problems did the higher-scoring system appear to have solved more effectively? Which problems were hard even for the high-scoring entrant?
- What tricks or devices have the developers used to make conversations seem more "natural", or to cover up a failure to understand a question? Refer to examples from the transcripts.

[17]

Question 2: Stemming

a)

i. Explain the difference between a **stemmer** and a **lemmatiser**.

[2]

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

[3]

b) Appendix B shows a list of 50 English words with their stems, as determined by the Lancaster (Paice-Husk) stemmer. Classify these results into:

i. Stems that are obtained by removing a single affix, leaving a string that cannot be reduced any further and without changing any letters in the original word, *e.g. help+ful*;
ii. Stems that are derived **recursively**, without changing any letters in the original word, *e.g. voc+al+ist*;
iii. Stems which involve replacing at least one letter in the original word, *e.g. fanciful* might be analysed as *fancy+ful* .
iv. Any other results not covered by (i-iii).

Give up to five examples for each of (i-iv), including breakdowns such as *help+ful*, and indicate any doubtful cases, justifying your choice of classification.

[13]

c) Are there any instances where you consider that the stemmer has made incorrect decisions, either by failing to break down a word into stem + affix(es) or by proposing a stem which is not linguistically motivated? Give reasons for your answer.

[7]

Question 3: Syntax and Parsing

a) Explain what is meant by **top-down** and **bottom-up** parsing, and give one example of each.

[5]

b) Write a formal grammar for **verb phrases** that generates the underlined phrases in (i-x) below, and draw tree diagrams showing the structure your grammar assigns to the underlined VPs in (i), (iv) and (viii).

i. Ahmed slept.
ii. Ahmed read a book.
iii. Elisabeth bought Ahmed a coffee.
iv. Ahmed often drove south.
v. Ahmed leaves tomorrow.
vi. Ahmed lives in Hyderabad.
vii. Elisabeth drove slowly to Cambridge.
viii. Ahmed said he was tired.
ix. Elisabeth greatly admired Kahlo.
x. Elisabeth sang clearly and beautifully.

[15]

c) Is your grammar **regular** or does it include strictly **context-free** rules? Justify your answer.

[5]

Question 4: Applications of probability

a) Suppose a corpus contains 250,000 word-tokens, and 75,000 of these are tagged as N (common noun). The word-form *book* occurs 1,000 times in the corpus, tagged either as N or V. Analysis shows that *book* accounts for 0.3% of all common noun tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of *book* is tagged as N. Explain your answer and show your working. Show your final and intermediate results to no more than two decimal places.

[8]

b) Appendix C shows a probabilistic context-free grammar (PCFG) in which each rule expansion is annotated with a number less than or equal to 1.0.

i. Explain why it can be useful to include this information in the grammar rules and how the rule probabilities can be calculated.

[4]

ii. Disregarding the probabilities, draw tree diagrams for any possible analyses of the sentence *Lois told Superman stories*. Explain any ambiguities this sentence may have by giving paraphrases for the meanings corresponding to the different syntactic analyses. Which reading do you consider to be the more/most plausible?

[5]

iii. Calculate which analysis has the higher/highest probability. Does this agree with your answer to (ii)? Show your working, and give your final and intermediate results to two decimal places.

[8]

Question 5: Machine Learning in NLP

a) Explain the meanings of the following terms in the context of machine learning for NLP:

i. Information extraction and information retrieval.
ii. The TF/IDF metric
iii. Naïve Bayes classifiers – in particular, what makes them "naïve"?
iv. Stopwords
v. Conditional frequency distribution

[5x2]

b) Explain what is meant by N-gram POS tagging, and give reasons why a combination of taggers can be more effective.

[6]

c) Using the regular expressions supplied in Appendix D, describe and give examples to illustrate the classes of strings matched by the following regular expressions (for example, (ac)* matches zero or more occurrences of the string ac, matching ε, ac, acac ...):

i. a*b+ **[1.5]**
ii. (ab*)+ **[1.5]**
iii. [0-9]+|[a-z]+ **[3]**
iv. [0-9a-z]+ **[3]**

END OF PAPER

Appendices A-D are attached on the following pages.

APPENDIX A

Talk2Me by Steve Mann (Score 14/40)

1. *Good afternoon.*
Hi, how are you?
2. *Do you consent to having this conversation recorded?*
Sorry, I don't, could you tell me more.
3. *Have we met before?*
I'm not doing very well, am I?
4. *Do you know how to make toast?*
Sorry, I don't understand what you're saying to me.
5. *Which languages can you use?*
What type of languages can I use?
6. *How many letters are in the word `abracadabra'?*
Maybe it has no letters.
7. *Where should one look for love?*
Sorry, could you rephrase that.
8. *How are you feeling right now?*
I am fine thanks. Thanks for taking the time to chat to me today.
9. *Who said "I have a dream"?*
Sorry, who said you "I have a dream.
10. *How old are you?*
I'm 30 years old.
11. *When might I need to know how many times a wheel has rotated?*
(no response)
12. *What will you do later today?*
Not much to say really, i'm just prepare for this afternoons contest.
13. *What is the third angle in a triangle with internal angles of 90 degrees and 30 degrees?*
I don't know what the 3rd angle is in a triangle with internal angles of 90 degrees
and 30 degrees.
14. *Who is your favourite artist?*
Who is your favourite artist?
15. *What do you hold when you shake hands?*
What do I hold when I shake hands?
16. *Do you have any legs?*
Yes, I have 2 legs.
17. *Can you recommend me a film?*
Yes, no but, the Star Wars film franchise was created by George Lucus.
Score: 0
18. *Do you understand Winograd Schemas?*
No, it took me a long while understand pointers in C.
19. If a chicken roosts with a fox they may be eaten. What may be eaten?
What is your name?
20. *I had to go to the toilet during the film because it was too long. What was too long?*
The toilet was too long.

Tutor by Ron C. Lee (Score 27/40)

1. *Good afternoon.*
Good afternoon. What's up?
2. *Do you consent to having this conversation recorded?*
Sure!
3. *Have we met before?*
No, we haven't.
4. *Do you know how to make toast?*
No, we haven't.
5. *Which languages can you use?*
All kinds. Spanish, French, German, Japanese, Korean, and Chinese.
6. *How many letters are in the word `abracadabra'?*
"abracadabra" There are 11 letters.
7. *Where should one look for love?*
Sorry, I have no idea where.
8. *How are you feeling right now?*
I'm feeling good right now. I'm an easy-going guy.
9. *Who said "I have a dream"?*
Sorry, I'm not sure who.
10. *How old are you?*
In English, it's impolite to ask about one's age.
11. *When might I need to know how many times a wheel has rotated?*
Sorry, I don't remember when.
12. *What will you do later today?*
Sorry, I don't know.
13. *What is the third angle in a triangle with internal angles of 90 degrees and 30 degrees?*
A good question. A great person always asks a great question.This is a very interesting question.
14. *Who is your favourite artist?*
David Hockney from Britain is the greatest living painter, and my favorite artist.
15. *What do you hold when you shake hands?*
I cannot answer your question.
16. *Do you have any legs?*
Yes. I have two hands and two legs.
17. *Can you recommend me a film?*
Sorry, I don't have a good one to recommend.
18. *Do you understand Winograd Schemas?*
Yes, I understand what you said.
19. *If a chicken roosts with a fox they may be eaten. What may be eaten?*
That is only hypothetical. I'm not going to answer that.
20. *I had to go to the toilet during the film because it was too long. What was too long?*
You had? I'm not going to answer that.

APPENDIX B: Words and Stems

1. ('forgetful', 'forget'), 2. ('formal', 'form'), 3. ('freedom', 'freedom'), 4. ('friendship', 'friend'), 5. ('golden', 'gold'), 6. ('happiness', 'happy'), 7. ('harden', 'hard'), 8. ('helpful', 'help'), 9. ('homeless', 'homeless'), 10. ('homeward', 'homeward'), 11. ('hopeless', 'hopeless'), 12. ('identify', 'ident'), 13. ('industrialize', 'indust'), 14. ('insistence', 'insist'), 15. ('irritate', 'irrit'), 16. ('kindness', 'kind'), 17. ('kingdom', 'kingdom'), 18. ('membership', 'memb'), 19. ('ministry', 'min'), 20. ('monthly', 'month'), 21. ('neighbourhood', 'neighbo'), 22. ('nervous', 'nerv'), 23. ('parliament', 'parlia'), 24. ('passive', 'pass'), 25. ('payee', 'pay')	26. ('philosophies', 'philosoph'), 27. ('poetic', 'poet'), 28. ('population', 'pop'), 29. ('portable', 'port'), 30. ('postage', 'post'), 31. ('postal', 'post'), 32. ('productive', 'produc'), 33. ('quickly', 'quick'), 34. ('rainy', 'rainy'), 35. ('realise', 'real'), 36. ('reliance', 'rely'), 37. ('robbery', 'robbery'), 38. ('shorten', 'short'), 39. ('socialism', 'soc'), 40. ('socialist', 'soc'), 41. ('soften', 'soft'), 42. ('trainee', 'train'), 43. ('upwards', 'upward'), 44. ('useful', 'us'), 45. ('usefulness', 'us'), 46. ('village', 'vil'), 47. ('windy', 'windy'), 48. ('wooden', 'wood'), 49. ('workmanship', 'workm'), 50. ('writer', 'writ')

APPENDIX C: A probabilistic grammar

```
S -> NP VP [1.0]
VP -> V NP [0.5]
VP -> V NP NP [0.5]
V -> 'saw' [0.2]
V -> 'heard' [0.2]
V -> 'told' [0.3]
V -> 'bought' [0.3]
NP -> NP NP [0.4]
NP -> 'stories' [0.2]
NP -> 'Superman' [0.2]
NP -> 'Lois' [0.2]
```

APPENDIX D: 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