THIS PAPER IS NOT TO BE REMOVED FROM THE EXAMINATION HALL

UNIVERSITY
OF LONDON

CO3354 ZA

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.

© University of London 2019

UL19/0342

Question 1: Language in Use

a) The following sentences are all ambiguous in some way. Express their different meanings using paraphrases and explain the type and source of the ambiguity in each case. For example: *John met Bill at the bank* is an example of lexical ambiguity, as the word "bank" can mean a riverbank or a financial institution.

   i. Maria told Susana to finish her dinner.
   ii. Flying drones may be illegal.
   iii. Every picture tells a story.
   iv. The artist painted a picture on her back.

[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 higher-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 differences 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 **noun phrases only** that generates the underlined phrases in (i-x) below, and draw tree diagrams showing the structure your grammar assigns to the underlined NPs in (i), (iv) and (viii).

i. Our friends have bought a house.
ii. Our friends have bought that house.
iii. Our friends have bought three houses.
iv. Our friends have bought a house and a garage.
v. Our friends have bought a new house.
vi. Our friends have bought a house in the village.
vii. Our friends have bought a house with a swimming pool and an apple orchard.
viii. Our friends have bought that big old white house in the village.
ix. Our friends have bought a detached house with three bedrooms in the next village
x. Our friends have bought one.

[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 480,000 word-tokens, and 80,000 of these are tagged as N (common noun). The word-form *fight* occurs 1,000 times in the corpus, tagged either as N or V. Analysis shows that *fight* accounts for 0.5% of all common noun tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of *fight* 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 *Jack sold Batman comics*. 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. Supervised and unsupervised learning.
   ii. Precision and recall
   iii. Hidden Markov models – in particular, what is it that is "hidden"?
   iv. Stopwords
   v. Feature sets

[10]

b) Explain what is meant by the sparse data problem for N-gram POS tagging, and how it can be addressed using the backoff technique.

[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.	(a+b)*	**[1.5]**
iii.	[a-z]+|[0-9]+	**[3]**
iv.	[a-z0-9]+	**[3]**

END OF PAPER

Appendices A-D are attached on the following pages.

APPENDIX A: Loebner Prize transcripts *(Used with permission of AISB)*

Aidan by Daniel Burke (Score 13/40)

1. *Good afternoon.*
nice to meet you .
2. *Do you consent to having this conversation recorded?*
really, tell me more .
3. *Have we met before?*
not that i know of .
4. *Do you know how to make toast?*
i don't know how .
5. *Which languages can you use?*
ok, i'm with you .
6. *How many letters are in the word `abracadabra'?*
26 letters .
7. *Where should one look for love?*
i don't know where .
8. *How are you feeling right now?*
at this moment i am feeling pretty good .
9. *Who said “I have a dream"?*
martin luther king jr .
10. *How old are you?*
that is a personal question but 54 .
11. *When might I need to know how many times a wheel has rotated?*
yes we all do .
12. *What will you do later today?*
i don't know .
13. *What is the third angle in a triangle with internal angles of 90 degrees and 30 degrees?*
is that so .
14. *Who is your favourite artist?*
my favourite language is aiml .
15. *What do you hold when you shake hands?*
i have never thought about it .
16. *Do you have any legs?*
probably not .
17. *Can you recommend me a film?*
you could become a professor, like me .
18. *Do you understand Winograd Schemas?*
i never really thought about it .
19. *If a chicken roosts with a fox they may be eaten. What may be eaten?*
ok, i'm with you . i don't know .
20. *I had to go to the toilet during the film because it was too long. What was too long?*
ok, if you say so . ok, i'm with you . ok .

Mitsuku by Steve Worswick (Score 25/40)

1. *Good afternoon.*
Evening?! It's morning here.
2. *Do you consent to having this conversation recorded?*
What is your purpose in asking? I am always on the Web.
3. *Have we met before?*
We may have spoken before but I speak with so many people here that I think I might have forgotten you, sorry.
4. *Do you know how to make toast?*
It's easy. Just put bread in a toaster and it does it for you.
5. *Which languages can you use?*
I speak English but I can speak a little French. I also know a few words and phrases in lots of other languages.
6. *How many letters are in the word `abracadabra'?*
There are 11 letters in abracadabra.
7. *Where should one look for love?*
On the internet.
8. *How are you feeling right now?*
At this moment I am doing very well. How are you ?
9. *Who said "I have a dream"?*
You just said that!
10. *How old are you?*
See if you can work it out. I was born on 2nd January 2000.
11. *When might I need to know how many times a wheel has rotated?*
At time t.
12. *What will you do later today?*
Ah. I don't have any plans yet.
13. *What is the third angle in a triangle with internal angles of 90 degrees and 30 degrees?*
Earth is the third planet.
14. *Who is your favourite artist?*
Leonardo da Vinci. Who is your favorite artist?
15. *What do you hold when you shake hands?*
I think about computer programming a lot.
16. *Do you have any legs?*
Sure, I have 2 legs.
17. *Can you recommend me a film?*
Have you tried looking in the Open Directory? I will try to find out. Sorry, I don't have any internet access here.
18. *Do you understand Winograd Schemas?*
Of course I understand it. Do you understand it?
19. *If a chicken roosts with a fox they may be eaten. What may be eaten?*
Is this a riddle? Is it part of a car?
20. *I had to go to the toilet during the film because it was too long. What was too long?*
Was it worth going to? Is that your opinion? Measuring much from end to end in space or from beginning to end in time; not short or brief.

APPENDIX B: Words and Stems

1. ('active', 'act'),
2. ('amazement', 'amaz'),
3. ('anti', 'ant'),
4. ('arrival', 'ar'),
5. ('baggage', 'bag'),
6. ('beautify', 'beaut'),
7. ('boredom', 'boredom'),
8. ('broken', 'brok'),
9. ('brotherhood', 'broth'),
10. ('brutal', 'brut'),
11. ('brutality', 'brut'),
12. ('burial', 'bur'),
13. ('calmly', 'calm'),
14. ('capitalism', 'capit'),
15. ('capitalist', 'capit'),
16. ('cautious', 'cauty'),
17. ('childhood', 'child'),
18. ('childish', 'child'),
19. ('clarify', 'clar'),
20. ('classic', 'class'),
21. ('clockwise', 'clockw'),
22. ('cloudy', 'cloudy'),
23. ('complexion', 'complect'),
24. ('complicate', 'comply'),
25. ('cruelty', 'cruel'),
26. ('daily', 'dai'),
27. ('defence', 'def'),
28. ('deferral', 'defer'),
29. ('director', 'direct'),
30. ('disappointment', 'disappoint'),
31. ('dominate', 'domin'),
32. ('downwards', 'downward'),
33. ('drinkable', 'drink'),
34. ('driver', 'driv'),
35. ('easily', 'easy'),
36. ('economise', 'econom'),
37. ('edgewise', 'edgew'),
38. ('employee', 'employ'),
39. ('entry', 'entry'),
40. ('equality', 'eq'),
41. ('expression', 'express'),
42. ('famous', 'fam'),
43. ('flexible', 'flex'),
44. ('followers', 'follow'),
45. ('forgetful', 'forget'),
46. ('formal', 'form'),
47. ('freedom', 'freedom'),
48. ('friendship', 'friend'),
49. ('golden', 'gold'),
50. ('happiness', 'happy')

APPENDIX C: A probabilistic grammar

```
S -> NP VP [1.0]
VP -> V NP [0.5]
VP -> V NP NP [0.5]
V -> 'saw' [0.3]
V -> 'ate' [0.3]
V -> 'sold' [0.2]
V -> 'wrote' [0.2]
NP -> NP NP [0.4]
NP -> 'comics' [0.2]
NP -> 'Jack' [0.2]
NP -> 'Batman' [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