Examiners' commentary 2018–2019

CO3354 Introduction to natural language processing – Zone B

General remarks

The examination was set as a mixture of questions that tested basic knowledge and understanding of the material ('bookwork'), and questions that require candidates to apply their knowledge and demonstrate deeper understanding by solving specific problems. There was a choice of three out of five questions, to enable candidates to address the areas in which they felt most confident. The examination paper also included an essay question, giving candidates an opportunity to show in-depth knowledge that may not have been covered by other questions and to develop an argument at length.

You are reminded to read each question carefully and address all aspects of the question. In particular, when instructed to 'explain' or 'describe' something, it is important to do so, rather than simply listing examples. As has been advised in previous years' commentaries, for any answers involving calculation, you should show your working, and worked examples are advisable in answers involving syntactic problems even if not specifically requested in the question. An incorrect solution may get credit where you are able to show some understanding of the problem through your working.

The cohort for this paper was quite small and not all questions were attempted. This commentary will therefore focus on explaining what would make good answers to the questions. All candidates obtained very good marks overall.

Comments on specific questions

Question 1: Language in use

a. Part (a) required candidates to identify the sources of various kinds of ambiguity including *referential*, *structural* and *lexical*. Ambiguity is pervasive in ordinary language and it poses challenges for NLP which practitioners need to be aware of. A useful exercise is to look out for ambiguous sentences or phrases when reading news reports – fairly soon you will spot them regularly.

b. Part (b) gave candidates an opportunity to reflect on some wider considerations about the purpose, achievements and 'state of the art' in NLP. This included the application of knowledge they may have acquired through self-directed readings, and evaluating the performance of 'chatbots', which ideally bring together a variety of language processing functions along with some general knowledge and reasoning capabilities. Good answers would note that in addition to grammatical and lexical knowledge, the dialogue required the chatbots to recognise and resolve instances of referential ambiguity between sentences, to test some areas of general knowledge and arithmetical reasoning, and to identify the use of particular strategies for masking the fact that a question had not been understood, some more convincing than others.

Question 2: Stemming

This question tested candidate's understanding of *stemming* and involved an evaluation of one particular approach to this task. This type of question comes up quite regularly in examinations and so candidates should be well prepared.

a. Part (a) was essentially bookwork, although appropriately selected examples from the text were required for full marks. Candidates were asked to explain the difference between two processes, a *stemmer* and a *lemmatiser*. In questions like this, it is important to give appropriate examples to avoid losing easy marks. The required definitions can be found in the appropriate parts of the subject guide and/or the essential textbook.

b. Part (b) involved analysing and evaluating the decisions made by a Lancaster (Paice-Husk) stemmer. Candidates had to decide whether stemming had been applied iteratively or in just one pass; an example of the former could be stemming *population* to *pop* by successively stripping *-ion, -at,* and *-ul*.

c. In (c), candidates were expected to explain *why* they judged that certain rules had been applied incorrectly, rather than to simply state that they were wrong. It is not enough to say that something can't be a stem because it is not a real word. An example could be reducing *ministry* to *min,* which suggests it may be related to *minor*, *minimum*, and so on.

Question 3: Syntax and parsing

This question concerned formal grammar and parsing. Some candidates had difficulty getting to grips with this kind of symbolic approach to language, and you are advised to revise the relevant sections of the subject guide and the recommended textbooks carefully. No candidates attempted this question.

a. Part (a) concerned a fundamental distinction between different classes of parsing regimes, *top-down* and *bottom-up*. This is an explicit learning outcome of Chapter 6 of the subject guide. Candidates who have completed this course are expected to have a good understanding of this distinction.

b. Part (b) involved proposing rules to match new data. For full marks, rules should be linguistically motivated rather than tweaked to generate particular strings: e.g. both nodes in a coordinate structure should be of the same phrasal category. Good answers would be both compact and generalisable, giving grammars that generate other grammatical sentences beyond the examples provided but do not allow for, or at least minimise, the production of ungrammatical sequences. Candidates often lose marks on questions of this kind by proposing 'flat', ad hoc rules that match the specific examples given but will not generalise to other grammatical sequences, or by neglecting to give tree diagrams as instructed.

c. Part (c) concerned the distinction between *context-free* and *regular* grammars. This distinction is fundamental in NLP and in CS in general, being closely related to automata theory. This topic often comes up in examinations, and many candidates seem to struggle with it. You are advised to revise it carefully, and to understand its significance by studying Chapter 2 of the subject guide and references given there.

Question 4: Applications of probability

This question concerned probabilistic methods in NLP.

a. Part (a) involved Bayes' Rule, which routinely comes up in examinations. As always, it is important to do all the question requires, including showing your intermediate calculations. You can still get some marks if you misremember the formula or if your calculations go awry, as long as some understanding is shown.

b. Part (b) concerned probabilistic grammars and parsing.

   i. The first sub-part was essentially book knowledge, covered in the subject guide and recommended textbooks.

   ii. In (ii) candidates should have found two different parses. For full marks, correct tree diagrams and clear paraphrases are required, along with good reasons for finding one of the analyses more plausible.

   iii. Part (iii) involved calculation of probabilities. Again, please bear in mind that you can gain some marks for explaining your answers and showing calculations, even if you do not end up with a correct solution. The question asked which analysis had the highest/higher probability. This could be answered correctly without including probabilities that are the same for all analyses, such as those for individual words. Candidates were also asked to compare the result with their intuitive understanding of the sentence. It should not be surprising if there is a difference, as the numbers have been arbitrarily specified for the purpose of this exercise, while interpretations are influenced by content and context as well as by prior expectations about the most likely structure.

Question 5: Machine learning in NLP

This question addressed some formal and mathematical techniques in NLP, including probabilistic reasoning and use of regular expressions.

i. Part (a) consisted of bookwork and should have been easily answerable by anyone who had read the subject guide and recommended textbooks thoroughly.

ii. Part (b) largely relied on book knowledge, as the answers can be found in section 4.5 of the subject guide. However, good answers will show not only recall of the content, but also the ability to explain the reasons for applying particular techniques.

iii. Part (c) involved regular expressions, an understanding of which is fundamental to many operations in NLP and computer science in general. For full marks in this kind of question it is important to make sure you answer the question properly and describe the expressions as well as giving examples.