Examiners' commentary 2018–2019

CO3354 Introduction to natural language processing – Zone A

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.

Comments on specific questions

Question 1: Language in use

This question was only answered by a few candidates, and most did not score high marks.

a. The first part of this question 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. Regrettably only one candidate did especially well on this part.

b. The second sub-question 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 you may have acquired through self-directed readings, by 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 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.

   Marks were awarded for showing appropriate technical knowledge, quality of argumentation, critical thinking and clarity of presentation.

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. Most candidates gave very good answers to this question.

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*. Most candidates did reasonably well. In questions like this, it is important to give appropriate examples as required to avoid losing easy marks. Some candidates disregarded this. The required definitions can be found in the appropriate parts of the subject guide and/or the recommended textbook.

b. Part (b) involved analysing and evaluating the decisions made by a Lancaster (Paice-Husk) stemmer. Most candidates who attempted this sub-question obtained very high marks, though a small number did not appear to understand what was required. Candidates had to decide whether stemming had applied iteratively or in just one pass. An example of the former could be stemming *locality* to *loc* by successively stripping *-ity* and *-al*.

c. In part (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 cannot be a stem because it is not a real word. An example could be reducing *anti* to *ant*, since there is no useful connection between these words.

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. However, most candidates did reasonably well on 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 in the subject guide. Some candidates obtained full marks for this sub-question but it was disappointing to see that a sizeable minority were not able to give good explanations.

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. For example, 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. The quality of answers varied noticeably. Quite a few were excellent and most were good or very good, though a minority showed poor understanding of the problem. Some candidates lost marks by proposing 'flat' ad hoc rules that matched the specific examples but would not generalise to other grammatical sequences.

c. Part (c) was the least well answered part of this question, though again some candidates provided excellent responses. The distinction between *context-free* and *regular* grammars is a topic that often comes up in examinations, and one that candidates often seem to struggle with, so you are advised to revise it carefully.

Question 4: Applications of probability

This question addressed some probabilistic techniques in NLP. This was the most successfully answered question in the paper, with the majority of candidates obtaining very good or excellent marks, while almost all who attempted it achieved at least good marks.

a. Part (a) involved Bayes' Rule, which routinely comes up in examinations. Most candidates seemed fairly confident with this and several obtained full marks. As always, it is important to do all the question requires, including showing your intermediate calculations. Note, the question asked for final and intermediate results to be shown to 'two decimal places', which the examiners treated as poor wording. No candidates were disadvantaged by this, and most answered to two significant figures, which is what was intended (this applies to part (b) as well). Remember to show your working, as you will get some credit 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, which is covered in the subject guide and recommended textbooks.

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

   iii. The final part (iii) involved the calculation of probabilities. Again, please bear in mind that you can gain some credit 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 was well answered by most candidates, though a few had rather poor marks.

a. Part (a) was essentially bookwork and should have been answerable by candidates who had read the subject guide carefully. In fact, marks were very variable, ranging from the maximum available to very low marks.

b. Part (b) largely relied on book knowledge, as the answers can be found in section 4.5 of the subject guide. However, good answers showed not only recall of the content, but also the ability to explain the reasons for applying particular techniques. The quality of answers was again very variable, ranging from full marks to zero or one of the six marks available.

c. Part (c) involved regular expressions, an understanding of which is fundamental to NLP and to computer science in general. It was gratifying to see that almost all answers were excellent, with quite a few candidates gaining full marks.