Examiners' commentary 2017–2018

CO3354 Introduction to natural language processing – Zone B

General remarks

The examinations were set generally as a mixture between questions that test 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, which should enable candidates to address the areas in which they feel most confident. Each paper 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, any answers involving calculation should show working, while worked examples are advisable in answers involving syntactic problems even if not specifically requested in the question. An incorrect solution can still get credit if 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 by candidates. This commentary will therefore focus on explaining what would make good answers to the questions..

Comments on specific questions

Question 1

This was an essay question, giving students an opportunity to reflect on some wider considerations about the purpose, achievements and state of the art in Natural Language Processing, and to apply knowledge they may have acquired through self-directed readings. This question referred directly to some points raised in the subject guide. For questions like this, marks are awarded for showing appropriate technical knowledge, quality of argumentation/critical thinking and clarity of presentation. To get good marks, it is important to structure your answer logically and make sure that the points you raise are relevant to the question. A frequent weakness encountered in answers to essay questions is that candidates do not draw sufficiently on specialist knowledge they should have acquired during the course, but show a level of expertise equivalent to that of a typical New Scientist reader.

Question 2

This question required candidates to show knowledge of formal grammars by applying a set of grammar rules and extending the rules to handle new data.

When devising new rules, it is important to aim for compactness and generality – your grammar should not be closely tailored to the specific examples, but should be capable of generating new grammatical sentences which are not in the example set, without also generating ungrammatical sequences. A useful

tip is to consider whether recursive rules are applicable (if you don't know what recursive rules are, make sure you do know before the examination).

Question 3

This question concerned probabilistic or weighted grammars.

Part (a) was essentially bookwork. Probabilistic grammars can be applied to disambiguation and gradient grammaticality: both of these are discussed in the NLTK book and the recommended text by Jurafsky and Martin, but only the former is discussed in the subject guide. They can also be incorporated into language models, which predict the probabilities of words in a sequence rather than structural parses.

Parts (b) and (c) involved identifying possible grammatical structures for a sentence and calculating their relative probabilities. Note that this can be done without calculating the "absolute" probabilities: you can disregard values that will be the same on any analysis, such as probabilities of individual words. However, you must show your calculations in detail including any intermediate results for full marks.

Question 4

This question involved stemming. As with the equivalent question in Zone A and similar questions in past years, there was a mixture of bookwork and problem solving: candidates were asked to hypothesise the rules that had been applied by the two stemmers and indicate where results may have been linguistically unfounded. It is important for your proposed rules to be as *general* as possible and to indicate how stemming decisions relate to different grammatical categories, such as plural forms and verb endings.

Question 5

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

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

Part (b) involved regular expressions, which are fundamental to many operations in NLP. 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.

Part (c) involved Bayes' Rule. Again, for full marks it is important to do all the question requires, including showing your intermediate calculations and the intended meanings of the variables in the formula. You will still get some credit if you misremember the formula or if your calculations go awry, as long as some understanding is shown.