Examiners’ commentaries 2015–16

CO3354 Introduction to natural language processing – Zone A

General remarks

The examination was set as a mixture of questions that test candidates’ 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, enabling candidates to address the areas in which they felt most confident. However, all questions involved problem-solving of some kind; candidates are unable to gain high marks on the basis of book-knowledge alone. Questions addressed both rule-based and data-driven approaches to natural language processing (NLP).

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. Candidates have been advised in previous years’ reports that any answers involving calculation should show working, and that worked examples are advisable in answers involving syntactic problems.

You must also take care to write legibly: you will only gain marks for answers that examiners are able to read.

The quality of answers was very good overall, with a high proportion of students achieving very good or excellent marks. There was some evidence that students had followed the advice in previous years’ reports on how to tackle particular questions.

Comments on specific questions

Question 1

This question addressed various aspects of empirical linguistics, including practical exercises and book-knowledge about corpus linguistics. Candidates who attempted this question mostly did reasonably well.

a. Answers were variable in quality. For full marks, candidates needed to give paraphrases bringing out the different meanings, and to show knowledge of different sources of ambiguity such as referential, logical scope, syntactic structure. Some candidates lost marks by failing to give clear paraphrases as required.

b. This was a straightforward ‘book-knowledge’ question, concerning terms which are explained in the subject guide. Most answers were reasonable but some showed little or no recall or understanding of these terms.

c. This question involved marking up a text with POS tags and discussing any decisions that had to be made for words that could have more than one tag. In general, candidates made a reasonable attempt at the annotation but provided limited discussion so lost marks, though some made sensible comments.

Question 2

This question tested candidates' ability to follow formal grammar rules and to formulate rules to cope with new data. Those who attempted this question generally did reasonably well.

a. Candidates were required to show understanding of formal grammar rules, including drawing syntax trees for specimen sentences. Most answers for this part were good and several candidates obtained full marks.

b. This question involved adding rules to match new data. Generally, candidates performed better on this question than part (a). The problem here required candidates to handle number agreement, which could be tackled in various different ways. Good answers used feature structures as described in section 6.3.2 of the subject guide, and included some sample rules and examples of worked derivations. For problems of this nature, you are advised to make proposed rules as concise and general as possible, rather than offering ad hoc solutions which are tailored to particular examples.

c. This question addressed a well-known problem with grammars that include recursive rules, which is discussed in the subject guide and has featured in previous exams. Marks were given for a clear and correct description of the problem, and for suggesting modified rules which would avoid it. Very few students did well on both, and a couple of answers seemed to miss the point entirely.

Question 3

This question addressed various formal and mathematical techniques in NLP, including probabilistic reasoning and use of regular expressions. Most answers were very good or excellent.

a. This was a 'book-knowledge' question: unusually, most candidates did slightly less well here than in the problem-solving parts (b) and (c), though a few achieved very high marks. All terms in this question are clearly defined in the subject guide and/or the NLTK book.

b. This question tested candidates' understanding and use of regular expressions: most answers were very good or excellent, with clear explanations and appropriate examples. Candidates seemed to be generally confident in using REs.

c. This part involved application of Bayes' rule, which is fundamental to computational linguistics (and AI in general). Most candidates did well and several obtained full marks, though not all provided a detailed explanation as required and some were confused. Generally, some credit would be given if students misremembered the rule but showed good working, or if their calculations went awry somewhere along the way. It is important to give clear explanations of your solution and show your calculations, so that examiners can award some marks.

Question 4

This question tested understanding of stemming and involved a comparison of two approaches to this task.

a. This question was essentially bookwork, though appropriately selected examples from the text were required for full marks. Most candidates did reasonably well, though surprisingly not as well as on the problem-solving part (b). In questions like this, candidates often lose marks by neglecting to give appropriate examples as required.

b. This part involved comparing and evaluating the decisions made by the two stemmers. Answers were mostly very good, with some candidates gaining full marks. Answers to questions of this type should be stated at as general a level as possible, rather than simply giving lists of words with and without their endings.

c. Candidates were expected to explain why they thought certain rules had been applied incorrectly, rather than simply stating that they were wrong. Generally, candidates performed noticeably less well on this part than the others.

Question 5

This question concerned probabilistic grammars and parsing. Relatively few candidates attempted this question but those who did achieved good to excellent marks.

a. This question was essentially book-knowledge. Probabilistic grammars can be applied to disambiguation and gradient grammaticality: both of these are covered in the recommended text by Jurafsky and Martin; only the former is discussed in the subject guide.

b. Candidates should have found three different parses. Correct tree diagrams and clear paraphrases were required for full marks.

c. This part involved calculation of probabilities; the quality of answers was extremely varied. Students in the final year of a computing programme should be confident with this kind of calculation, but this appears to be an area of weakness for some. You should bear in mind that you may get some credit for explaining your answers and showing calculations, even if you do not end up with a correct solution.

Examiners’ commentaries 2015–16

CO3354 Introduction to natural language processing – Zone B

General remarks

The examination was set as a mixture of questions that test candidates’ 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, enabling candidates to address the areas in which they felt most confident. However, all questions involved problem-solving of some kind; candidates were unable to gain high marks on the basis of book-knowledge alone. ‘Basic knowledge’ is not limited to the content of the subject guide; candidates are expected to have extended their knowledge by studying the NLTK book and other resources. Questions addressed both rule-based and data-driven approaches to natural language processing.

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. Candidates have been advised in previous years’ reports that any answers involving calculation should show working, and that worked examples are advisable in answers involving syntactic problems.

You must also take care to write legibly: you will only gain marks for answers that examiners are able to read.

The cohort in Zone B was rather small this year so it is not possible to generalise from their results, which ranged from very good to excellent. This report focuses on the requirements for good answers, rather than analysing candidates’ performance.

Comments on specific questions

Question 1

This question addressed various aspects of empirical linguistics, including practical exercises and book-knowledge about corpus linguistics.

a. Answers should show knowledge of different sources of ambiguity such as referential, logical scope, syntactic structure. For full marks, clear and accurate paraphrases are required.

b. This question is a straightforward book-knowledge question about terminology in corpus linguistics: all terms are explained in chapter 3 of the subject guide. Clear, correct and concise definitions are required for full marks.

c. This question involved marking up a text with POS tags and discussing any decisions that had to be made for words that could have more than one tag. In other exams with similar questions, candidates have tended to make a reasonable attempt at the annotation but have provided limited discussion. It is important to look out for words which can occur as, for example, verbs or nouns ('close', 'father', 'play', 'state'), and to explain how the context determines which category is chosen.

Question 2

This question tests the ability to follow formal grammar rules and to formulate rules to cope with new data.

a. It is important to keep in mind that the sentences asked for in (a. (i)) must be grammatical – not just licensed by this grammar, which also allows for ungrammatical sequences.

b. This question involved adding rules to match new data. Rules are required to cover different verb categories, such as transitive and intransitive. For full marks, you should explain the issue, and give some sample rules and worked derivations. Proposed new or modified rules should aim for generality, and not just handle the specific examples in the question. You should also be sure to code lexical rules for any words which are not in the original grammar.

c. Candidates were required to extend the grammar further to deal with adjectives and adverbs. As with part (b), answers should aim for generality and compactness; this may involve recursive rules. For example, as well as 'a big old wooden desk' the grammar should handle 'a big old red wooden desk' and so on. Marks were awarded for correctly identifying the problem, proposing appropriate modifications to the grammar and giving worked examples.

Question 3

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

a. This was a book-knowledge question, involving some terminology from quantitative empirical linguistics. All the terms in this question are clearly defined in the subject guide and/or the NLTK book.

b. This question tested understanding and use of regular expressions. This question should be straightforward given that all RE operators are documented in an appendix. Answers showed that candidates were generally confident in the use of REs.

c. This question involved application of Bayes' rule, which is fundamental to modern computational linguistics and AI in general. Good answers showed familiarity with the rule and ability to carry out probabilistic calculations. The latter should be well within the competence of final-year computing undergraduates. For full marks, it is important to show working and to explain your answer. You will be given credit for this even if you end up with an incorrect answer.

Question 4

This question tested understanding of stemming and involved a comparison of two approaches to this task.

a. This question was bookwork, involving terms which are explained in the subject guide and the NLTK book.

b. This question involved comparing and evaluating the decisions made by the two stemmers. Answers to questions of this type should be stated at as general a level as possible, rather than simply lists of words with and without various affixes. It is important to make explicit comparisons between the outputs of the two stemmers, rather than simply listing them and leaving the reader to draw their own conclusions. For instance, the

Lancaster stemmer is generally considered to be more ‘aggressive’ than Porter and answers should discuss whether this is borne out by the examples provided in this question.

c. Candidates were expected to judge and explain whether certain rules had been applied incorrectly, and to justify their claims rather than simply stating that the results were wrong. For example, the Lancaster stemmer has removed the ending –th from south, as if it were an ordinal term like fourth, fifth.

Question 5

This question concerned probabilistic grammars and parsing.

a. This question is essentially book-knowledge. The distinctions between a grammar and a parser, and top-down vs bottom-up parsing, are explained in the subject guide and the NLTK book.

b. Candidates should have found three different parses. Correct tree diagrams and clear, accurate paraphrases were required for full marks.

c. This question involved calculation of probabilities. Students in the final year of a computing programme should be confident with this kind of calculation. As with question 3, it is important to show working and give clear explanations. The question also required candidates to discuss whether the resulting probabilities accord with their intuitive understanding of the sentence. Good answers noted that preferred readings depend on context and semantic content as well as syntactic probability.