Examiners' commentaries 2016–17

CO3354 Introduction to Natural Language Processing – Zone A

General remarks

The examination was set generally as a mixture between questions that test your basic knowledge and understanding of the material ('bookwork'); and questions that require you to apply your knowledge and demonstrate deeper understanding by solving specific problems. There was a choice of three out of five questions, which should enable you to address the areas in which you feel most confident. However, all questions involved problem-solving of some kind; you would be 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.

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' examiner commentaries, any answers involving calculation should show working; and 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 quality of answers was good overall, with a high proportion of students achieving very good or excellent marks.

Comments on specific questions

Question 1

This question tested the ability to follow formal grammar rules and to formulate rules to cope with new data. Those who attempted this question generally did reasonably well overall, although there was a noticeable variation in marks.

Part (a) required candidates to show understanding of formal grammar rules, including drawing syntax trees for sentences. The standard of answers varied quite widely: some students achieved full marks, but others did less well, either by composing sentences which did not match the grammar or by failing to give all allowable structures for their constructed sentences.

Part (b) involved adding rules to match new data. Generally, candidates did better on this than on part (a), although again there was a noticeable variation in standards. The problem here was to handle **coordination** of words and phrases and **iteration** of adjectives. For problems of this nature, you are advised to make your proposed rules as concise and general as possible, rather than offering ad hoc solutions which are tailored to particular examples. In the case of coordinating rules of the form $A \rightarrow B\ Conj\ C$, the symbols A, B and C should all be of the same syntactic category. Analysing, *e.g.* example (i) as *S and V* is not a good solution, as it is too closely tailored to this particular sentence. A rule which licenses sequences like *a difficult long monograph* should allow for phrases with

an indefinitely long sequence of adjectives rather than precisely two. Note that the question asks for 'appropriate worked examples': you should not expect the examiners to laboriously satisfy themselves that the rules match the specimen sentences.

Part (c) addressed a well-known problem with grammars that include recursive rules, which is discussed in the subject guide and has regularly featured in past examinations. Marks were given for a clear and correct description of the problem, and for suggesting modified rules which would avoid it. Very few candidates did well on both parts here, and several did not even attempt this sub-question. Note that the question specifically asks for suggested modifications to the grammar rules, so answers which involve using a different parsing regime would be missing the point.

Question 2

This question concerned probabilistic grammars and parsing. The standard of answers was generally higher than in previous years, indicating that students had put some effort into preparing for this type of question.

Part (a) is essentially book knowledge. 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. Chapter 8, section 6.3 in the NLTK book gives a rather terse explanation of these two terms, as well as an example of a weighted grammar and a sample execution using a Viterbi parser from the toolkit.

Part (b): you should have found three different parses. Correct tree diagrams and clear paraphrases were required for full marks.

Part (c) involved calculation of probabilities. You should bear in mind that you can gain credit for explaining your answers and showing calculations, even if you do not end up with a correct solution. Question 2c(i) asked which analysis had the highest probability: this could be answered correctly without including probabilities which are the same for all analyses, such as those for individual words. Question 2c(ii) asked candidates 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, and interpretations are influenced by content and context as well as by prior expectations about the most likely structure.

Question 3

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

Part (a) is essentially bookwork, although appropriately selected examples from the text were required for full marks. Most candidates did reasonably well, generally better than on the problem-solving sub-question (b). In questions like this, it is important to give appropriate examples as required to avoid losing easy marks.

Part (b) 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 giving lists of words with and without their endings. It is important to address all parts of the question, including the requirement to discuss motivations for the rules: typically, this would involve considerations of whether particular word-endings can be reliably associated with grammatical forms, and whether rules appear to apply iteratively. An example of the latter could be stemming *removes* to *remov* by successively stripping *-s* and *-e*.

In (c), you were expected to explain **why** you judged that certain rules had been applied incorrectly, rather than to simply state that they were wrong. One example would be stripping *-th* from the end of words as if they were ordinals, which is not appropriate for forms like *mouth*, *Wentworth*. Generally, candidates performed noticeably less well on this part than on the others.

Question 4

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.

Part (a) answers were mostly good but rather 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 lexical, referential, syntactic structure (*e.g.* PP-attachment). Some candidates lost marks by failing to give clear paraphrases as required.

Parts (b) and (c) are straightforward book knowledge questions, again with some variation in the quality of answers. Those who have thoroughly read and assimilated the subject guide and supplementary reading should obtain good marks on this kind of question.

Part (d) 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. Candidates often lose marks on this kind of question by neglecting the second point: this text has several examples of words that can belong to more than one category such as *top* (noun, verb, adjective …) *musical* (noun, adjective). It is sensible to track these decisions while marking up the text. Good answers explained the reasons for tagging decisions in terms of, for example, context and/or semantic content.

Question 5

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.

Part (a) is a book knowledge question, although terms like *constituent structure*, *N-gram tagging*, *tokenisation* refer to concepts that you need to understand in order to tackle problem-solving tasks. All the terms in this question are clearly defined in the subject guide and/or the NLTK book. The standard of answers noticeably ranged from excellent to rather weak.

Part (b) tests understanding and use of regular expressions (REs): most answers were very good or excellent, with clear explanations and appropriate examples. A few candidates showed limited confidence in tackling this question. Given that finite-state methods in general are fundamental in many areas of NLP, and the course has covered various applications of REs such as POS-tagging and stemming, it is important to have a sound understanding of these concepts.

Part (c) involves application of Bayes' Rule, which is fundamental to modern 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 got into a muddle. Generally, some credit was given if candidates 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, to be sure of picking up marks.

Examiners' commentaries 2016–17

CO3354 Introduction to Natural Language Processing – Zone B

General remarks

The examination was set generally as a mixture between questions that test your basic knowledge and understanding of the material ('bookwork'), and questions that require you to apply your knowledge and demonstrate deeper understanding by solving specific problems. There was a choice of three out of five questions, which should enable you to address the areas in which you feel most confident. However, all questions involved problem-solving of some kind; you would be unable to gain high marks on the basis of book knowledge alone. 'Basic knowledge' is not limited to the content of the subject guide; you are expected to have extended your knowledge by studying the NLTK book and other recommended readings, together with sources identified through independent study. 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. As has been advised in previous years' examiner commentaries, any answers involving calculation should show working; and worked examples are advisable in answers involving syntactic problems even if not specifically requested in the question.

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

Comments on specific questions

Question 1

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

Part (a): it is important to keep in mind that the sentences asked for in (a) i. should be grammatical – not simply licensed by this grammar, which also allows for ungrammatical sequences.

Part (b) involved adding and/or modifying rules to match new data and to prevent over-generation. 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. You should not expect examiners to laboriously satisfy themselves that the grammar matches the data. Proposed new or modified rules should aim for concision and generality rather than being tailored to the specific examples in the question.

Part (c) addressed a well-known problem with grammars that include recursive rules, which is discussed in the subject guide, and has regularly featured in

past examinations. Marks were given for a clear and correct description of the problem, and for suggesting modified rules which would avoid it. Note that the question specifically asks for suggested modifications to the grammar rules, so answers which involve using a different parsing regime would be missing the point.

Question 2

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

Part (a) is 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.

Part (b) tested understanding and use of regular expressions. This question is straightforward as all RE operators are documented in an appendix. Answers showed that candidates were generally confident in the use of REs.

Part (c) involved application of Bayes' Rule – $P(A|B) = P(B|A)P(A)/P(B)$ – which is fundamental to modern computational linguistics and AI in general. Good answers would show familiarity with the Rule and ability to carry out probabilistic calculations. For full marks, it is important to show working and provide some explanation, making clear which condition corresponds to A and which to B in the above formulation, and showing each step in the calculation. You will be given credit for doing this even if you end up with an incorrect answer, but show some understanding of the problem.

Question 3

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

Part (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.

Parts (b–c) involved straightforward 'book knowledge' about terminology in corpus linguistics: all terms are explained in the subject guide. Clear, correct and concise definitions are required for full marks. Those students who have carefully read and assimilated the guide will be well prepared for questions like this.

Part (d) 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. Candidates have tended to make a reasonable job of the annotation but have skimped on the discussion. It is important to look out for words which can occur as, for example, either verbs or nouns *(`will', 'build', 'group')*, and to explain how the context determines which category is chosen.

Question 4

This question concerned probabilistic grammars and parsing.

Part (a) is essentially book knowledge. 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. Chapter 8, section 6.3 in the NLTK book gives a rather terse explanation of these two terms, as well as an example of a weighted grammar and a sample execution using a Viterbi parser from the toolkit.

Part (b): candidates should have found three different parses. Correct tree diagrams and clear, accurate paraphrases were required for full marks.

Part (c) involves calculation of probabilities. It is important to show working and give clear explanations, as this enables examiners to give credit for partial understanding in the event of incorrect solutions. The question also requires discussion of whether the resulting probabilities accord with the candidate's intuitive understanding of the sentence. Good answers would note that preferred readings depend on context and semantic content as well as syntactic probability.

Question 5

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

Part (a) is bookwork, involving terms which are explained in the subject guide and the NLTK book.

Part (b) involved comparing and evaluating the decisions made by the two stemmers (Porter and Lancaster). 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 various affixes. Good answers might discuss whether particular word endings are associated with grammatical categories, and whether rules are applied iteratively: for instance, *conducted* may be reduced to *conduc* by successively stripping *-ed* and *-t*. It is important to make explicit comparisons between the outputs of the two stemmers, rather than simply listing them and leaving the examiners to draw their own conclusions. For instance, the Lancaster stemmer is generally considered to be more 'aggressive' than Porter and answers could discuss whether this is borne out by the examples provided in this question.

In (c), candidates are expected to explain if they judge that certain rules have been applied incorrectly. It is important that you justify these claims rather than simply stating that the results are wrong. For example, the Lancaster stemmer has removed the ending *–th* from *earth,* as if it were an ordinal term like *fourth, fifth*, and both tools have stemmed some proper names. Lancaster, but not Porter, has truncated Roman names, which is questionable given that its rules are designed for the English language.