Exam May 2015, Questions

Natural Language Processing (University of Sheffield)

The University Of Sheffield.

Data Provided:
None

DEPARTMENT OF COMPUTER SCIENCE **Spring Semester 2014-2015**

NATURAL LANGUAGE PROCESSING **2 hours**

Answer THREE questions.

All questions carry equal weight. Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

1. a) Explain why smoothing is important for the practical application of N-gram language models. Illustrate your answer with a suitable example. Briefly describe the approach behind Good-Turing discounting. [30%]

   b) Hidden Markov Modelling (HMM) is a popular approach to automatic part-of-speech tagging. A bigram HMM tagger's estimate of the best tag sequence t_1^n for a given word sequence w_1^n is expressed by the formula:

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}}\ P(t_1^n|w_1^n)$$

   where this is approximated by assuming:

$$\underset{t_1^n}{\operatorname{argmax}}\ P(t_1^n|w_1^n) \approx \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

   Explain these formulae, making it clear how the approximation is derived and what the two simplifying assumptions are that underlie it. [30%]

   c) Consider the sentence *We conduct research*. Since *conduct* and *research* are both nouns as well as verbs this sentence can be tagged in at least four ways.

   The following counts are observed in a corpus: (a) unigram word/tag counts; (b) bigram tag counts; (c) counts of occurrences of a word with a particular part of speech tag. (Here <s> denotes a special start of sentence marker.)

	Count
We	7342
conduct	45
research	253
<s>	50000
VB	148787
NN	253048
PRP	63520

(a)

	VB	NN	PRP
<s>	789	5783	2304
VB	43	7432	1038
NN	1134	2276	1358
PRP	1492	68	9

(b)

	We	conduct	research
VB	0	25	81
NN	0	30	172
PRP	63520	0	0

(c)

   Formally define, as a probabilistic finite state machine, a minimal HMM capable of tagging the example sentence. You do not need to calculate all of the transition probabilities or the observation likelihoods, but you should show how you would calculate at least one of each. (Note: If you do not have a calculator you may express these probabilities in the form of fractions.)

   [40%]

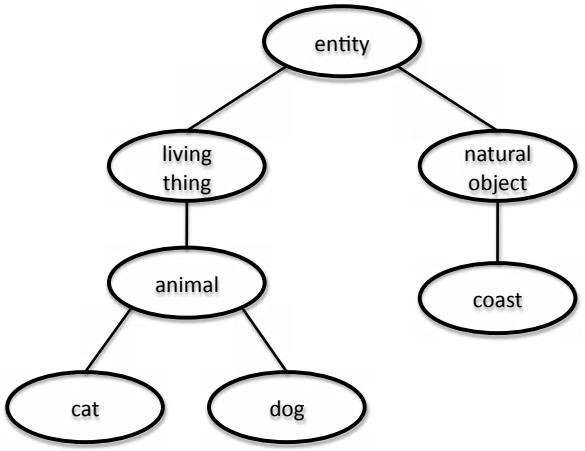
2. a) Describe the problem of Word Sense Disambiguation. Explain the differences between supervised, semi-supervised and dictionary approaches to the problem and give an example of each approach. [30%]

b) Explain what is meant by the problem of computing word similarity. Briefly describe the two main approaches that have been applied to this problem. [20%]

c) Information Content based measures of word similarity use corpus counts to estimate the probability of concepts in a hierarchy. Show how these values are calculated and how Information Content values can be derived from them, given the counts observed from a corpus and fragment of a thesaurus hierarchy shown below. (Note: If you do not have a calculator you may leave your answers in the form of arithmetic expressions involving logarithms.)

	Count
animal	2
cat	10
coast	5
dog	8

(a): Corpus counts



(b): Fragment of thesaurus hierarchy

[30%]

d) Using the three measures of semantic similarity that are based on Information Content, determine the similarity scores that would be generated by each for the following pairs of concepts based on the probabilities computed in 2)c). (Note: If you do not have a calculator you may leave your answers in the form of arithmetic expressions involving logarithms.)

(i) *cat* and *dog* [10%]

(ii) *cat* and *coast* [10%]

3. a) The following sentences each contain an instance of a natural language 'generalised quantifier' expression (shown underlined). For each such expression, show how its meaning can be characterised by a formula that includes a condition on sets, in accordance with Generalised Quantifier Theory.

   (i) *Two artists visit Montmartre.*

   (ii) *John keeps no more than three fish.*

   (iii) *Few men who are healthy smoke.* [10%]

   b) Formal semantics is based on two assumptions, known as the *principle of compositionality* and the *rule-to-rule hypothesis*.

   (i) Describe the principle of compositionality and explain its relevance to the observation that language use is highly productive. [10%]

   (ii) Explain the rule-to-rule hypothesis and characterise its relation to the principle of compositionality. [10%]

   (iii) Illustrate your answers to 3(b)(i) and 3(b)(ii) by giving a formal syntactic and semantic analysis of a simple example phrase or sentence. [10%]

   c) Imagine that we want to write a grammar which restricts the sentences it generates to allow only *semantically plausible* phrases to fill the roles of subject and object. To simplify, we will allow only the transitive verbs *eats* and *rides*, and the intransitive verb *falls*, and likewise only the nouns *boy*, *dog*, *sandwich* and *bike*. These nouns have various attributes that make them suitable or not suitable as either the subject or object of the different verbs (excluding metaphoric uses). For example, the subject of *eats* should be *animate* (as is *boy* or *dog*, but not *sandwich* or *bike*), whilst its object should be *edible* (like *sandwich*, but not *bike*). We will assume that the subject of *rides* must be *human* (here only *boy*), but allow that the subject of *falls* is unrestricted. Therefore, some of the sentences that the grammar should admit (✓) or exclude (×) are as follows:

Sentence	*Accepted?*
the boy eats the sandwich	✓
the boy eats the bike	×
the dog eats the sandwich	✓
the bike eats the sandwich	×
the boy rides the bike	✓
the dog rides the bike	×
the sandwich rides the bike	×
the boy falls	✓
the sandwich falls	✓

Specify a grammar that uses feature representations (including features for the attributes *human, animate* and *edible*) to encode the restrictions described above. The grammar should generate the sentences in the above table that are marked as "✓", and not those marked as "×". Your grammar should cover both lexical entries and phrase structure rules. You should use the PATR notation for writing feature-based grammars in stating your answer. [30%]

d) Show how the syntax of English relative clauses can be represented using a phrase-structure approach that includes slash categories (or uses a slash feature). As part of your answer, specify a grammar that is sufficient for analysing the relative clause *which Billy owns*, and draw the tree structure that your grammar assigns to it. [30%]

4. Consider the following example sentence: *I shot the elephant in my pyjamas.*

This sentence is syntactically and semantically ambiguous, in that the prepositional phrase *in my pyjamas* may modify either the preceding noun or the verb.

a) Define a grammar which uses simple atomic categories (i.e. *not* feature representations) and allows alternative analyses of the above sentence corresponding to the two alternative meanings. Draw the trees your grammar allows for the example. [20%]

b) Describe the parsing strategy of a *recursive descent parser*. What are the disadvantages of this parsing approach? Illustrate your answer with reference to the example sentence given above. [35%]

c) Describe an algorithm for bottom-up chart parsing. Your answer should include an account of the data structures used. How does this chart parsing algorithm address the problems of simple recursive descent parsing? Illustrate your answer with reference to the example sentence given above. [45%]

END OF QUESTION PAPER