COMP34411

Two hours

Question ONE is COMPULSORY

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Natural Language Systems

Date: Thursday 21st January 2016

Time: 09:45 - 11:45

Please answer Question ONE in Section A and TWO Questions from Section B.

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

[PTO]

Section A

You should answer question 1: each part of this question carries 5 marks

1. a) Explain, with examples, the difference between lexical, structural and scope ambiguity. **[3 marks]** Which of these is the most significant problem for machine translation systems (and say why)? **[2 marks]**

   b) What are precision, recall and F-measure? **[3 marks]**. Suppose you had a system that prescribed drugs in response to natural language descriptions of symptoms. Would it be better for this system to have high precision or high recall (and say why)? **[2 marks]**

   c) What is the difference between the way that people use their articulators to produce vowels and the way they use them to produce consonants? **[3 marks]** On the basis of what you have just said about this difference, decide which of the following speech signals and intensity contours corresponds to someone saying '*kick*' and which to someone saying '*icky*'. **[2 marks]**

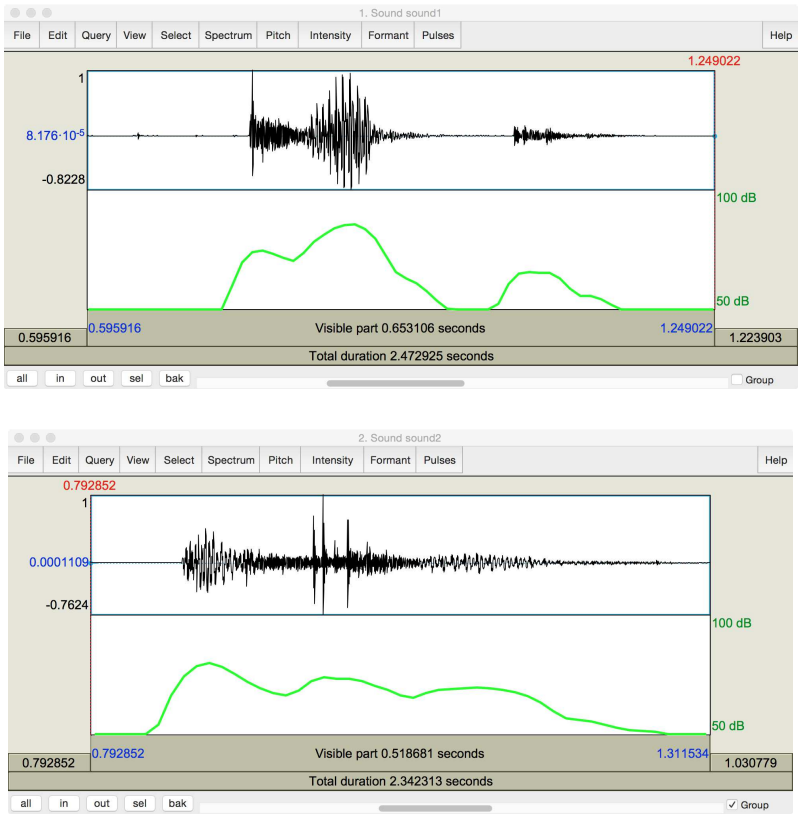


Figure 1: Recordings of '*kick*' and '*icky*'

   d) What are 'transfer rules'? Illustrate your answer by sketching a rule for turning

English sentences (where the normal word order is subject-verb-object) to Martian sentences (where the normal word order is object-subject-verb). **[4 marks]**. Martians distinguish between things which are biddle and things which are boddle, and this is marked on the form of NPs in Martian. Would it be easier to write transfer rules that turn English NPs into Martian ones or *vice versa*? **[1 mark]**

e) What is a 'vector space' model of meaning? **[3 marks]**. You should illustrate your answer by showing simple versions of the representations of the following three passages:

(a) Manuel Pellegrini says he is delighted with the team for coming from behind to win on the night.

(b) try these quick and easy squid recipes from authentic calamari to the simple grilled squid in lemon and garlic.

(c) Andalucian paella: an authentic seafood paella, boasting some of the finest ingredients, from calasparra rice to chorizo.

On the basis of the information available in these three sentences, it is impossible to tell which ones are most closely related. Where might you find the information that you would need in order to realise that (b) and (c) are more closely related to each other than either is to (a)? **[2 marks]**

f) Suppose you have been training a part-of-speech tagger on subsets of some large corpus, and the curve relating accuracy to the size of the training corpus is as shown in Fig. 2. Sketch what you would expect the curve to look like as the size of the training data increased to 2500. **[2 marks]**. Explain why it is important to separate training data from testing data, and describe what is mean by 'N-fold cross validation'. When might doing 2-fold cross validation be a good idea, and when might using 20 folds be a good idea? **[3 marks]**

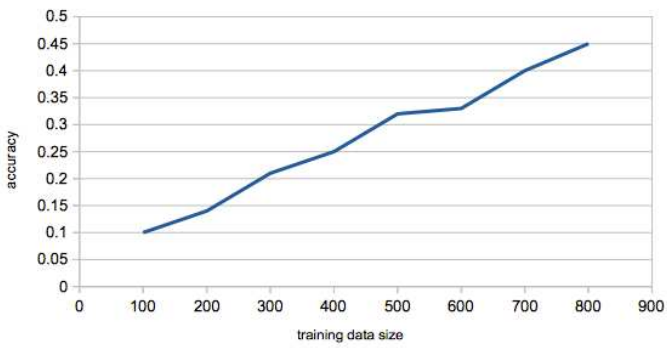


Figure 2: Accuracy against training data size for a part-of-speech tagger

TOTAL **[30 marks]**

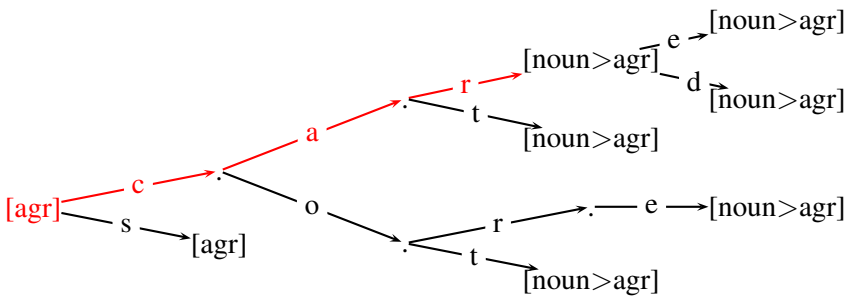
Section B

You should answer two questions from this section: each question carries 35 marks

2. a) Describe the use of 'categorial' descriptions of the items that combine together to make words. **[5 marks]** Suppose that Martian contains words of type scrinch, where a scrinch takes a following scronch and a preceding scranch, e.g. that there is a word '*bidtongin*' which is made out of the root '*tong*', the prefix '*bid*' and the suffix '*in*'. Give a categorial description of these three morphemes. **[6 marks]**

b) Outline an algorithm for looking up words where the morphemes of the language in question are stored in a letter trie, there are a set of spelling rules that specify what can happen at the boundaries between morphemes, and morphemes are given categorial descriptions the specify how they can combine. **[8 marks]**

Suppose you are using the algorithm you have described to look up the word '*cares*' in the lexical trie shown in Fig. 3, with the spelling rules given in this figure, and it has followed the arcs marked in red in this figure. What would it do for its next four moves? **[7 marks]**



Unseen: es, items found so far: [noun>agr]

```
[v0] ==> [e, +, v0] : [v1, c0] _ [c1];
```

Figure 3: Part way through looking up '*cares*'

c) Explain how 'Brill retagging' can be used to improve the performance of simple-minded taggers, e.g. frequency-based approaches. Illustrate your answer using the tagged text and the templates in Fig. 4. **[9 marks]**

Tagged text

	she	knows	that	the	man	knows	that	she	loves	him
Gold standard	PRON	VB	COMP	DET	NN	VB	COMP	PRON	VB	PRON
Tagger output	PRON	VB	PRON	DET	NN	VB	PRON	PRON	VB	PRON

Templates

```
#t1(T1, T2, T3, T4): T1 > T2 if tag[0]=T3 and tag[1]=T4;
#t2(T1, T2, T3, T4): T1 > T2 if tag[0]=T3 and tag[-1]=T4;
```

Figure 4: Text and templates for Brill retagging

3. a) Nivre has argued that there are three measures that you might want a parsing algorithm to optimise, namely accuracy, speed and robustness. Describe a situation in which accuracy matters more than speed and one where speed matters more than accuracy. **[5 marks]**

   b) Describe the major data structures and operations used in the 'arc-eager' version of the MALT parser **[7 marks]**. Explain why this algorithm is guaranteed to have the best time complexity that any parser can have. **[3 marks]**

   c) Outline how you could use a dependency treebank to train a machine learning algorithm to learn which of the operations used by the algorithm to apply in a given setting. **[10 marks]**

   d) Martian has two part of speech tags, `brev` and `norp`, and allows words to appear in absolutely any order. Show that it is not possible for the algorithm you have described above to assign the dependency relations implied by the tree in Fig. 5 for the sentence '*ehs tahw swonk eh sekil*'. **[10 marks]**

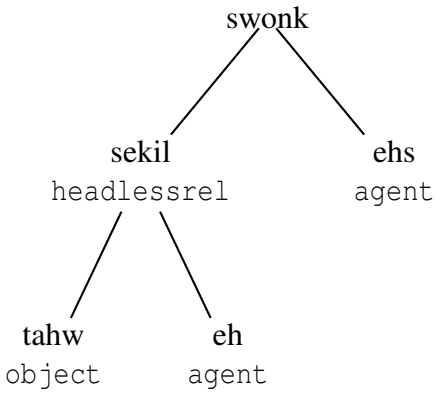


Figure 5: Dependency tree for '*ehs tahw swonk eh sekil*'

4. a) Describe the difference between textual entailment and strict entailment. **[2 marks]** Give a situation in which textual entailment would be more useful than strict entailment and one where strict entailment would be more useful than textual entailment. **[3 marks]**

b) Describe the dynamic time warping algorithm, and explain why it is generally sensible for the cost of an exchange to be lower than the sum of the costs of an insert followed by a delete. **[6 marks]** Given costs of 3 for an exchange and 2 for insert and delete, say what the cost of aligning '*whey*' and '*way*' is. **[4 marks]** The matrix in Fig. 6 shows an intermediate state of the algorithm: you may use this as the basis of your calculation.

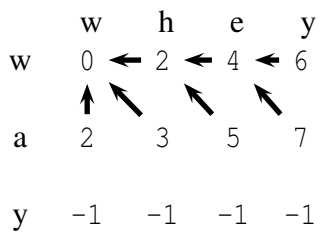


Figure 6: Dynamic time warping applied to '*whey*' and '*way*'

c) How could you use this algorithm for determining textual entailment? Give an example where doing this will produce a strict entailment and one where it will produce something which is not strictly valid. **[6 marks]** What would you have to do to exploit the WordNet subset/superset relations in this application? **[3 marks]** What would happen if you applied this to the pairs in (1) and (2)? **[2 marks]**

(1) TEXT. John loves Mary
HYPOTHESIS. John likes Mary

(2) TEXT. John doesn't love Mary
HYPOTHESIS. John doesn't like Mary

d) Approximate matching of dependency trees generally gives more reliable results than string matching. Outline an approximate matching algorithm which would enable you, given the trees in Fig. 7, to infer that '*John saw a policeman in the park*' entails '*John saw a man*'. **[6 marks]** Describe the complexity of this algorithm. **[3 marks]**

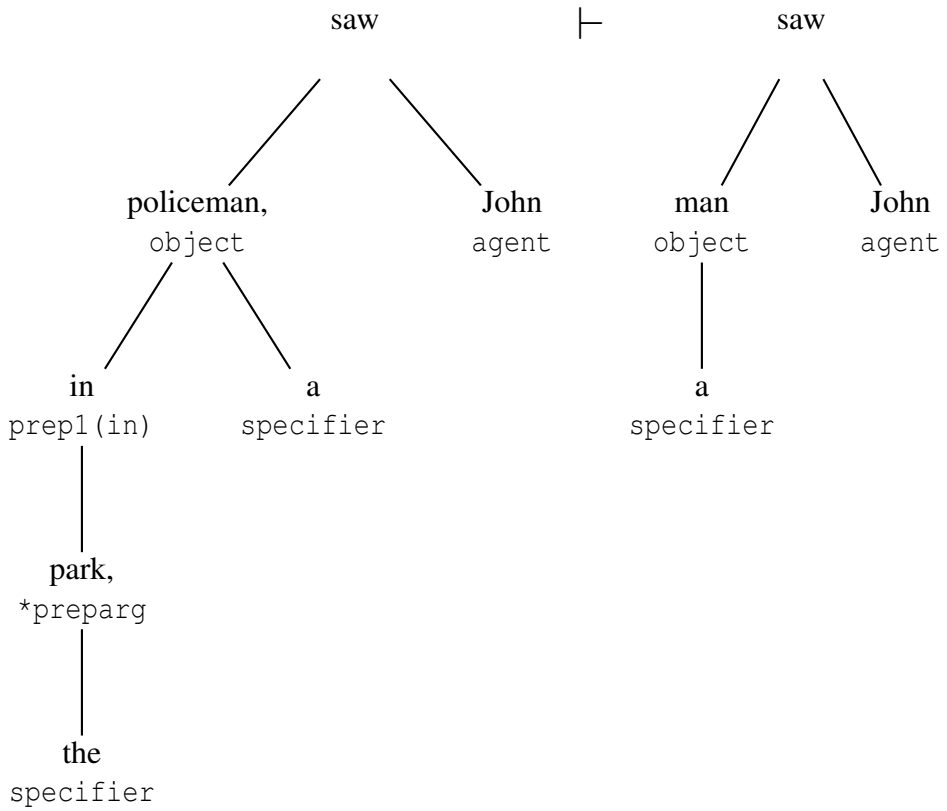


Figure 7: Does '*John saw a policeman in the park*' entail '*John saw a man*'?

5. a) What are 'transition probabilities' and 'emission probabilities', and how are they used within hidden Markov models (HMMs)? **[8 marks]** Use the HMM in Fig. 8 to estimate the most likely part-of-speech tags for the sentence '*He runs races*'. **[7 marks]**

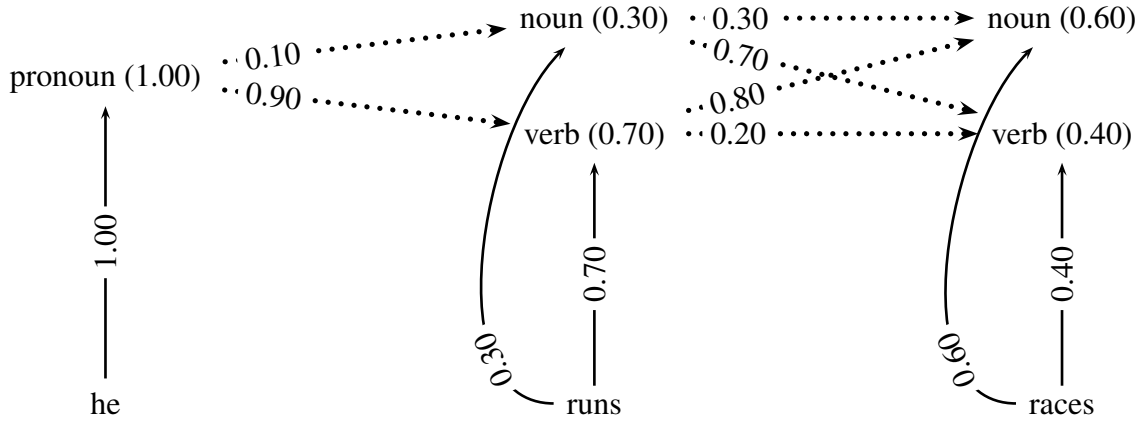


Figure 8: HMM for part-of-speech tagging

b) The training phase for the HTK estimates the emission and transition probabilities for a collection of HMMs, which are then connected together during the recognition phase to make a single large HMM for recognising complete utterances. Before this can be done you have to specify a context free grammar: what is the role of the grammar? You should illustrate your answer with reference to the grammar and lexicon given in Fig. 9. **[8 marks]**

```
$s = $np $verb;
$np = he | she;
$verb = runs | sleeps;

he = h iy
she = sh iy
runs = r u n s
sleeps = s l iy p s
```

Figure 9: Grammar and lexicon for speech recognition

c) Describe how diphone-based speech synthesis works. **[4 marks]** For diphone synthesis to work well, you have to obtain a set of diphones which have been extracted from a body of recorded material. Explain how a speech recogniser can be used to find the boundaries of diphones in recorded speech. **[4 marks]** Suppose you had a recording of someone saying '*He entered the room*': which of the diphones that you could extract from this recording would be useful if you wanted to synthesise '*She inferred the answer*'? **[4 marks]**

END OF EXAMINATION