

## 684.02 Assignment 9: Short questions

Chris Brew

March 11, 2002

1. You are called as a neuropsychological expert witness in a court case. An unsigned typewritten confession has been found at the scene of the crime. You have examined the confession and concluded that it was written suffering from a minor linguistic impairment called Herzog's aphasia<sup>1</sup>. About 1% of the population suffer from this syndrome. Medical records reveal that the defendant has the disease.

The prosecution argue that "There is a 1% chance that the defendant would have the disease if he was innocent, so there is a 99% chance that he committed the crime". On the other hand the defence argue that "There are 100,000 people in this town, so about 1,000 of them have the disease, so there is only a 1 in 1000 chance that the defendant is guilty."

Both prosecution and defence are (perhaps deliberately) making mistakes about conditional probabilities. Briefly explain to the judge why neither defence nor prosecution should be trusted in this matter.

(15 marks for each problem, total marks 30)

2. Describe the differences between:
  - supervised and unsupervised training
  - part-of-speech tagging and shallow parsing.
  - probabilistic context-free grammar and lexical dependency grammar
  - Hidden Markov Models and Markov Models

---

<sup>1</sup>The name is imaginary: you shouldn't look it up.

Word	Non-uniform model	Uniform model
the	0.375	0.2
cat	0.125	0.2
dog	0.25	0.2
chases	0.125	0.2
sleeps	0.125	0.2

Table 1: Probabilities for a unigram model of English

- (1) the cat chases the dog
- (2) the dog sleeps
- (3) the cat sleeps
- (4) the dog chases the dog

Table 2: A tiny corpus of English

Briefly illustrate the differences with reference to tasks and/or in natural language processing. (6 marks per distinction, total of 24 marks)

3. Consider two unigram models of (a tiny part of) English in which only five words are possible. In these models the emission probability for a word does not depend on its context. The probabilities for two models (called *uniform* and *non-uniform*) are listed in table 1. A (similarly tiny) corpus of sentences is shown in table 2. For each model:
  - (a) Calculate the probability for each sentence in the corpus under this model. (4 marks per model)
  - (b) Calculate the probability of the whole corpus under this model. (1 mark per model)
  - (c) Using the formula:

$$H = - \sum_i p(w_i) \log(p_M(w_i)) \quad (5)$$

calculate the cross-entropy  $H(M)$  of the model with respect to the corpus. Note that  $p(w_i)$  (the true probability) is got by counting words in the corpus, and that  $p_M(w_i)$  is read off directly from the provided model. (6 marks per model)

Why is the cross-entropy larger for the uniform model than for the non-uniform model? (4 marks)

(Total of 26 marks)

4. This question is about the performance of natural language systems.
  - (a) Why is it much harder to achieve 95% accuracy in statistical parsing than in part-of-speech tagging. (5 marks)
  - (b) How well do you expect a human being to do on these tasks? How would you evaluate whether the system is performing as well as a human being? (5 Marks)

(total of 10 marks)

5. Explain the difference between a monitor corpus and a reference corpus. (10 marks)