2011 Paper 7 Question 9

Natural Language Processing

(*a*) Give an equation for finding the most probable sequence of part of speech (POS) tags that could be utilised by a stochastic POS tagger. You should assume a bigram model. [5 marks]

(*b*) Given the following training data, show the estimates that would be obtained for the probabilities in the equation you gave:

```
the_DT0 green_AJ0 bottle_NN1 leaked_VVD ._PUN the_DT0
suppliers_NN2 bottle_VVB water_NN1 ._PUN green_AJ0 water_NN1
suppliers_NN2 bottle_VVB ._PUN
```

[4 marks]

(*c*) Explain what is meant by the terms *smoothing* and *backoff* in the context of stochastic POS tagging. [4 marks]

(*d*) One common source of errors in stochastic POS taggers is that nouns occurring immediately before other nouns (e.g. *catamaran trailer*) are often tagged as adjectives and, conversely, prenominal adjectives are often tagged as nouns (e.g. *trial offer*). Suggest possible reasons for this effect. [7 marks]