2005 Paper 8 Question 14

Natural Language Processing

In (1) and (2) below, the words in the sentences have been assigned tags from the CLAWS 5 tagset by a stochastic part-of-speech (POS) tagger:

(1) Turkey_NP0 will_VM0 keep_VVI for_PRP several_DT0 days_NN2 in_PRP a_AT0 fridge_NN1

(2) We_PNP have_VHB hope_VVB that_CJT the_AT0 next_ORD year_NN1 will_VM0 be_VBI peaceful_AJ0

In sentence (1), *Turkey* is tagged as a proper noun (NP0), but should have been tagged as a singular noun (NN1). In sentence (2), *hope* is tagged as the base form of a verb (VVB: i.e., the present tense form other than for third person singular), but should be NN1. All other tags are correct.

(a) Describe how the probabilities of the tags are estimated in a basic stochastic POS tagger. [7 marks]

(b) Explain how the probability estimates from the training data could have resulted in the tagging errors seen in (1) and (2). [6 marks]

(c) In what ways can better probability estimates be obtained to improve the accuracy of the basic POS tagger you described in part (a)? For each improvement you mention, explain whether you might expect it to improve performance on examples (1) and (2). [7 marks]