COMPUTER SCIENCE TRIPOS Part IB, Part II 50% − 2022 − Paper 7

5 Formal Models of Language (apc38)

This question concerns the noisy channel framework for problem-solving in natural language processing.

(a) You have the following sentences translated from English into Triposi.

English	Triposi
she drinks water	mwamni sileng
the rain soaks the teacher	sileng mworob sesesrakan
the teacher drinks here	sesesrakan mwamni mwabma
the teacher keeps drinking	sesesrakan mwatbo mwamni

  (i) Translate the following English sentences into Triposi:
      the rain keeps soaking the teacher
      she keeps drinking water here   [2 marks]

  (ii) Describe how we can calculate the likelihood of a translation using the noisy channel framework: you will need to give and explain the equation for decoding from one language into another, and explain how you can obtain the information needed to carry out the calculations.   [4 marks]

  (iii) What problem do the following underlined English words present, given the training data we have so far, and how can you still translate them into Triposi with a machine?
      the <u>teachers</u> *drink* <u>Perrier</u>   [3 marks]

(b) You know that your optical character recognition model has made errors on every predicted instance of 'e'. Explain what information you need in order to automatically correct the errors with a noisy channel approach.   [3 marks]

(c) A signaller sends you Morse code messages, but you know that they send a dot when they should send a dash two times in five, and a dash instead of a dot three times in ten. You also know that they use the character M $(- -)$ 3 times in 100, N $(- \cdot)$ and I $(\cdot \cdot)$ 7/100, and A $(\cdot -)$ 8/100.

   You receive the message " $\cdot \cdot$ $- \cdot$ $\cdot -$ ". What is the likelihood that it represents MIN, MAN, NAN or AIM?   [4 marks]

(d) Discuss the similarities and differences between the noisy channel and human processing of spoken English.   [4 marks]