5 Formal Models of Language (pjb48)

Alice and Bob have both installed an app that uses a language model exclusively trained on examples of their conversations with each other. In this app, Bob can watch words appear on his device's screen as Alice types into her device.

(*a*) Today, Alice and Bob are arranging to meet. Alice is writing to Bob. Bob watches Alice's words appear on his screen. The words say *I'll see you at the* and then she gets interrupted. The language model's three most probable continuations are *lab*, *canteen*, and *plodge*. Calculate the surprisal (in bits) of all three given that: $P(lab \mid \text{context}) = 0.5$, $P(canteen \mid \text{context}) = 0.125$, $P(plodge \mid \text{context}) = 0.03125$. Provide relevant equations. [3 marks]

(*b*) At noon the next day, Bob is arranging to meet Alice for lunch. He is in a rush and messages her *c u @ cnTn*. Alice understands the message to mean *see you at the canteen*.

(*i*) With reference to information theory, explain how Alice was able to decode the message. [5 marks]

(*ii*) Alice sends a voice note reply. She says: *No, I'll meet you at, uhm, well, the chronophage*. With reference to ideas from information theory, explain why the inclusion of these fillers might help Bob process the high-surprisal word *chronophage*? [3 marks]

(*c*) Alice and Bob discover that the model assigns high surprisal to the word *arrived* in the sentence *The lecturer who the students like arrived late*. Given that their model is an n-gram model, explain why this might be. [3 marks]

(*d*) Alice wants to reduce the bandwidth used by the app. She proposes a probabilistic encoding scheme where the number of bits used to transmit a word is proportional to its surprisal.

(*i*) Discuss the efficiency of Alice's scheme with reference to the sentence *The lecturer who the students like arrived late*. [3 marks]

(*ii*) Make suggestions of ways to improve the language model to reduce bandwidth further, commenting on feasibility. [3 marks]