Formal Models of Language

Paula Buttery

Easter 2020

The following comprises suggested questions for discussion in Supervision 2 (lectures 5-8). Supervisors understand that it's exam term—be prepared to discuss the ideas if you don't have time to tackle all the questions.

Natural Languages

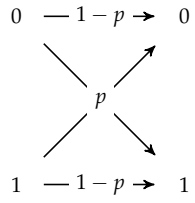
1. Write a paragraph on the following open-ended point to discuss with your supervisor:
   - Is frequency information or structural information more important when considering processing difficulty?
2. Give examples and counter-examples of sentences in English (or any other language) that would support theories of constant information rate (try to think of examples that are different from those provided in the slides!).

Formal Languages and Learnability

1. Consider a rigid classic categorial grammar $C_{cg} = (\Sigma, Pr, S, \mathcal{R})$ where $Pr = \{S, X\}$, $\Sigma = \{a, b, c, d, e\}$ and S=S. If a, b have type X and you know that $abc \in \mathcal{L}(C_{cg})$, $abdc \in \mathcal{L}(C_{cg})$, $ebc \in \mathcal{L}(C_{cg})$, give possible types for each of c, d and e?
2. Explain why a finite class of finite languages is learnable within Gold's paradigm.
3. Describe a learning paradigm where a learner could learn from two sources simultaneously (a bilingual)

Information Theory

1. A binary symmetric channel is one where the input x_i and the output y_i are both in $\{0, 1\}$. The channel is characterised by p the probability that an input bit is transmitted as the opposite bit. If q is the probability that the source sends $x = 0$, and $1 - q$ the probability of $x = 1$, show that the mutual information is maximised when zeros and ones are transmitted with equal probability (i.e. when $q = 0.5$).



2. Using the processed Alice in Wonderland file (course materials page) write some simple code to generate some good candidates for nonsense words by:
   - finding the probability distribution defined by a bigram language model
   - generating some words using the probability distribution
   - selecting the 10 words whose information rate is lowest
3. Describe how you could frame the following tasks as noisy channel problems:
   - automatically answering questions
   - disambiguating multiple senses of a word

Distributional Models

1. Describe how you might use word distributions to compare the similarity of two characters in a text. What might any *similarity* be telling us about the characters?