Ling 684.01: Lecture Notes 15
From Incremental Recognition to Cognitive Modeling

1. Prefix (forward) probabilities in human processing problems:

   Some related work to information-theoretic processing models:

   - [Shannon, 1948] – quantify info in communication as bits, given distribution
   - [Hale, 2001] – surprisal: elim prob mass costs time to ramp up lower prob hypoths
   - [Just and Varma, 2007] – FMRI: ambiguity leads to recruitment of other areas
   - [Levy, 2008] – particle filter model: quanta of prob mass
   - [Wu et al., 2010] – PPDA (HHMM) prefix probs correlate with time delays
   - [Hale, 2006] – entropy reduction: delays correspond to sharpening distribs

2. Information theory:

   Goal was to optimize signal through noisy channel (phone line)

   Formalize communication:

   transmit propositions x over which is known distribution $\mathsf{P}(x)$

   (if distribution is not known, assume non-optimal uniform distrib)

   More efficient to encode common propositions in fewer bits

   if uniform:

event/prop	freq.	code	cost
cd	25	00	50b
ls	25	01	50b
mv	25	10	50b
rm	25	11	50b
	100		200b

   if not uniform:

event/prop	freq.	code	cost
cd	50.0	00	100b
ls	25.0	01	50b
mv	12.5	10	25b
rm	12.5	11	25b
	100		200b

   more optimal code:

event/prop	freq.	code	cost
cd	50.0	0	50.0b
ls	25.0	10	50.0b
mv	12.5	110	37.5b
rm	12.5	111	37.5b
	100		175.0b

more information is inherent in more peaked distribution, requires fewer bits

Entropy expressed in bits, logarithmic on size of domain

Entropy shows number of bits needed to send message (as fn of expected distrib):

$$\begin{aligned}
\mathsf{H}(x) &= \lim_{N\to\infty} -\frac{1}{N}\log_2(\mathsf{P}(x_1)\cdot\mathsf{P}(x_2)\cdot\mathsf{P}(x_3)\cdot\dots\cdot\mathsf{P}(x_N)) \\
&= \lim_{N\to\infty} -\frac{1}{N}\log_2(\prod_{x\in X}\mathsf{P}(x)^{\mathsf{P}(x)\cdot N}) \\
&= \lim_{N\to\infty} -\frac{1}{N}\sum_{x\in X}\mathsf{P}(x)\cdot N\cdot\log_2\mathsf{P}(x) \\
&= -\sum_{x\in X}\mathsf{P}(x)\log_2\mathsf{P}(x) \qquad \text{(average bits over } N\text{)}
\end{aligned}$$

Opposite of predictability:

fair coin:	$\mathsf{H}(x) =$	$-.5\log_2 .5 - .5\log_2 .5$	$=$	$.5 + .5$	$= 1$
one-side hemisphere:	$\mathsf{H}(x) =$	$-.99\log_2 .99 - .01\log_2 .01$	$=$	$.01 + .07$	$= .08$

3. Cognitive relevance:

Information concept used to formulate surprisal, a way to quantify change in a distribution:

$$\log_2\frac{\mathsf{P}(x_1,\dots,x_t)}{\mathsf{P}(x_1,\dots,x_{t-1})} = \log_2\frac{\sum_{y_t}\mathsf{P}(y_t,x_1,\dots,x_t)}{\sum_{y_{t-1}}\mathsf{P}(y_{t-1},x_1,\dots,x_{t-1})}$$

Relation to cognitive model:

- Neurons activate/support hypotheses in proportion to probability.
- When best hypotheses $\mathsf{P}(x_t \mid \dots) = 0$, neurons take time to reallocate activation.
- Finite supply of neurons; when there aren't enough stay active, reader 'garden paths.'

References

[Hale, 2001] Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166, Pittsburgh, PA.

[Hale, 2006] Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):609–642.

[Just and Varma, 2007] Just, M. A. and Varma, S. (2007). The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, & Behavioral Neuroscience*, 7:153–191.

[Levy, 2008] Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

[Wu et al., 2010] Wu, S., Bachrach, A., Cardenas, C., and Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 1189–1198.