CS/IS 6742: NLP and Social Interaction, Fall 2017

Nov. 2, 2017: lecture 21, modeling (local) language structure: “phrase” space, heading towards intra-sentential structure

---

1 “Trailer”

Two papers it might be worth skimming over the next few days or the next week:

Hale, John. 2001. A probabilistic earley parser as a psycholinguistic model. Proceedings of NAACL, pp. 1-8.

Levy, Roger and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Proceedings of NIPS, 849-856.

I mention these papers now because the topic connects to:

- our discussion last lecture of the constant entropy principle (Genzel and Charniak, 2002), also known as the uniform information density principle (Levy and Jaeger, 2007)
- upcoming discussion of intra-sentential syntactic structure
- the notion that language is social, involving people communicating

2 The Brown clustering n-gram language model

$$P(w_k|w_1^{k-1}) = P(w_k|c_k)P(c_k|c_1^{k-1}) \tag{1}$$

3 Useful information-theoretic quantities

3.1 Reminders

Surprisal:

$$-\log(r_i) = \log \frac{1}{r_i} \tag{2}$$

Surprisal can also be considered to be "amount of information", although to some the intuition seems backwards. An analogy: suppose you know that an event e happens with probability 1. Then e happens. Have you learned anything from e happening? No; so you have gained no information from it.

If we consider the "reference" distribution to be q, then the *cross-entropy*

$$H(q||r) = \sum_i q_i \log \frac{1}{r_i} \tag{3}$$

is the expected surprisal for r with respect to reference distribution q.

The Kullback-Leibler (KL) divergence is a "corrected" cross-entropy achieving a minimum of 0 at $q = r$:

$$D(q||r) = \sum_i q_i \log \frac{q_i}{r_i} \tag{4}$$

3.2 "Derived" quantities

The *entropy* (think of it as the "self cross-entropy"):

$$H(q) = \sum_i q_i \log \frac{1}{q_i} \tag{5}$$

The *mutual information* can be considered to be the KL divergence between the joint distribution of two random variables and the joint distribution *if they were independent*.

We exemplify in terms of the Brown clustering paper. Let us suppose that our random variables are C_1 and C_2, meaning something like "the next cluster (or word type)" and "the cluster (or word type) that would immediately follow". Then, we can consider the KL divergence between $P_{\text{dependent}} = p(C_1, C_2)$ and $P_{\text{independent}} = p(C_1)p(C_2)$:

$$\sum_{c_1,c_1} p(C_1, C_2) \log \frac{p(C_1, C_2)}{p(C_1)p(C_2)} = \sum_{c_1,c_1} p(C_1, C_2) \log \frac{p(C_2|C_1)}{p(C_2)} \tag{6}$$