Supplementary Material C

Formal derivation of information locality

The aim of this section is to show that, under progressive erasure noise, processing cost increases when context words that predict a target word are distant from that target word. A derivation of this result was originally presented in Futrell and Levy (2017). The simpler derivation here was originally presented in Futrell (2019).

Assume that the memory encoding function M is structured such that some proportion of the information available in a word is lost depending on how long the word has been in memory. For a word which has been in memory for one timestep, the proportion of information which is lost is a constant e_1; for a word which has been in memory for two timesteps, the proportion of information lost is e_2; in general for a word which has been in memory for t timesteps, the proportion of information lost is e_t. Assume further that e_t is monotonically increasing in t: i.e. $t < \tau$ implies $e_t \leq e_\tau$. This memory model is equivalent to assuming that the context is subject to erasure noise, where the erasure rate is assumed to increase with time, a noise distribution we call **progressive erasure noise**.

Under progressive erasure noise, the memory representation r of the context $w_1, \ldots, w_{i-1}$ can be represented as a sequence of symbols $r_1, \ldots, r_{i-1}$. Each symbol r_j, called a **memory symbol**, is equal either to the context word w_j or to the erasure symbol $\mathtt{E}$. The surprisal of a word w_i given the memory representation $r_1, \ldots, r_{i-1}$ can be written in two terms:

$$-\log p(w_i | r_1, \ldots, r_{i-1}) = -\log p(w_i) - \text{pmi}(w_i; r_1, \ldots, r_{i-1}),$$

where $\text{pmi}(w_i; r_1, \ldots, r_{i-1}) = \log \frac{p(w_i|r_1,\ldots,r_{i-1})}{p(w_i)}$ is the **pointwise mutual information** (Church & Hanks, 1990; Fano, 1961) of the word and the memory representation, giving the extent to which the particular memory representation predicts the particular word. We can now use the chain rule to break the pointwise mutual information into separate terms, one for each symbol in the memory representation:

$$\begin{aligned}\text{pmi}(w_i; r_1, \ldots, r_{i-1}) &= \sum_{j=1}^{i-1} \text{pmi}(w_i; r_j | r_1, \ldots, r_{j-1}) \\ &= \sum_{j=1}^{i-1} \text{pmi}(w_i; r_j) - \sum_{j=1}^{i-1} \text{pmi}(w_i; r_j; r_1, \ldots, r_{j-1}) \\ &= \sum_{j=1}^{i-1} \text{pmi}(w_i; r_j) - R, \end{aligned} \tag{21}$$

where pmi$(x; y; z)$ is the three-way pointwise **interaction information** of three variables (Bell, 2003), indicating the extent to which the conditional pmi$(w_i; r_j | r_1, \ldots, r_{j-1})$ differs from the unconditional pmi$(w_i; r_j)$. These higher-order interaction terms are then grouped together in a term called R.

Now substituting Eq. 21 into Eq. 3 (repeated below), we get an expression for processing difficulty in terms of the pmi of each memory symbol with the current word:

$$D_{\text{lc surprisal}}(w_i | w_1, \ldots, w_{i-1}) \propto \mathop{\mathbb{E}}_{r | w_1, \ldots, w_{i-1}} [-\log p(w_i | r)] \tag{3}$$

$$\begin{aligned} &= \mathop{\mathbb{E}}_{r | w_1, \ldots, w_{i-1}} [-\log p(w_i) - \text{pmi}(w_i; r)] \\ &= \mathop{\mathbb{E}}_{r | w_1, \ldots, w_{i-1}} \left[-\log p(w_i) - \sum_{j=1}^{i-1} \text{pmi}(w_i; r_j) + R \right] \\ &= -\log p(w_i) - \mathop{\mathbb{E}}_{r | w_1, \ldots, w_{i-1}} \left[\sum_{j=1}^{i-1} \text{pmi}(w_i; r_j) + R \right] \\ &= -\log p(w_i) - \sum_{j=1}^{i-1} \mathop{\mathbb{E}}_{r_j | w_j} [\text{pmi}(w_i; r_j)] + \mathop{\mathbb{E}}_{r | w_1, \ldots, w_{i-1}} [R]. \end{aligned} \tag{22}$$

It remains to calculate the expected pmi of the current word and a memory symbol given the distribution of possible memory symbols. Recall that each r_j is either equal to the erasure symbol $\texttt{E}$ (with probability e_{i-j}) or to the word w_j (with probability $1 - e_{i-j}$). If $r_j = \texttt{E}$, then pmi$(w_i; r_j) = 0$; otherwise pmi$(w_i; r_j) =$ pmi$(w_i; w_j)$. Therefore the expected pmi between a word w_i and a memory symbol r_j is $(1 - e_{i-j})$pmi$(w_i; w_j)$. The effect of erasure noise in the higher-order terms collected in R is more complicated, but in general will have the effect of reducing their magnitude, because a higher-order interaction information term will have a value of 0 whenever any single variable in it is erased. Therefore we can write the expected processing difficulty per word as:

$$D_{\text{lc surprisal}}(w_i | w_1, \ldots, w_{i-1}) \propto -\log p(w_i) - \sum_{j=1}^{i-1} (1 - e_{i-j}) \text{pmi}(w_i; w_j) + o(K), \tag{23}$$

where $o(K)$ indicates a value that is is bounded by K, and K is the sum of all higher-order interaction information terms involving the words $w_1, \ldots, w_{i-1}$.

Next, we subtract the value of $D_{\text{surprisal}}$ from $D_{\text{lc surprisal}}$ to get an expression for **memory distortion** (introduced in Supplementary Material A), which is the excess processing cost induced by memory limitations, above and beyond the processing cost predicted by plain surprisal theory. Assuming the higher-order terms collected in $o(K)$ can be neglected, the memory distortion comes out to:

$$D_{\text{lc surprisal}}(w_i|w_1, \ldots, w_{i-1}) - D_{\text{surprisal}}(w_i|w_1, \ldots, w_{i-1}) = \sum_{j=1}^{i-1} e_{i-j}\text{pmi}(w_i; w_j), \tag{12}$$

which was the expression given in Section 5.1. As words w_i and w_j become more distant from each other, the value of the erasure probability e_{i-j} must increase, so the value of Eq. 12 must increase. Therefore the theory predicts increased processing difficulty as an increasing function of the distance between w_i and w_j in direct proportion to the pointwise mutual information between them.

If we include the effects of the higher-order terms collected in K, then Eq. 23 also implies that processing difficulty will increase when groups of elements with high interaction information are separated from each other in time. See Bell (2003) for the relevant technical details on interaction information.