Transition Freedom; Peak Freedom

An attempt to define these two quantities, please offer corrections if this is wrong. – Linas 1 July 2022

Formal Defintion

Let t be a token drawn from a vocabulary $T = \{t\}$ of size $|T|$. Let $w = t_1 t_2 \cdots t_n$ be an n-gram. Then, given the observed sequence (w,t) of an n-gram followed by a 1-gram, define the following:

- Let $N(w,t)$ be the number of times that the sequence (w,t) was observed.
- Let $N(w,*) = \sum_t N(w,t)$ be the sum over counts of all such sequences.
- Let $\Delta(w,t) = \begin{cases} 1 & \text{if } N(w,t) > 0 \\ 0 & \text{if } N(w,t) = 0 \end{cases}$ be the "Dirac delta" or "indicator function".
- Let $\Delta(w,*) = \sum_t \Delta(w,t)$ be called the "transition freedom" (I think this is the correct defintion of transition freedom, is that correct?)

The forward "peak freedom" is then defined as

$$\Delta(t_1 t_2 \cdots t_n, *) - \Delta(t_2 t_3 \cdots t_{n+1}, *)$$

is that correct?

The reverse peak freedom is then

$$\Delta(*, t_1 t_2 \cdots t_n) - \Delta(*, t_2 t_3 \cdots t_{n+1})$$

Is that right, or am I off-by-one in this defintion?

Other norms are

- Let $N_p(w,*) = \sum_t N^p(w,t)$ be the power norm (like the ℓ_p norm but without the root).
- Clearly $\Delta(w,*) = N_p(w,*)\big|_{p=0}$ is just the limit.
- Let $S(w,*) = \frac{1}{\log 2} \cdot \frac{d}{dp} N_p(w,*)\Big|_{p=0} = \sum_t \log_2 N(w,t)$ be an entropy.
- Let $H(w,*) = \frac{1}{\log 2} \cdot \frac{d}{dp} N_p(w,*)\Big|_{p=1} = \sum_t N(w,t) \log_2 N(w,t)$ be a weighted entropy.

The two entropy variants $S(w,*)$ and $H(w,*)$ are interesting, as they minimze the contribution of stray, accidental markup. That is, if $N(w,*)$ is a million, and there's a stray t such that $\Delta(w,t) = 1$, then $\Delta(w,*)$ is larger by one, than it would otherwise be. Meanwhile, both $S(w,t)$ and $H(w,t)$ are unchanged.