W. ROSS MORROW

1. Introduction

Consider a corpus $\mathcal{C}$ of tokens from a language $\mathcal{T} = \{1, \dots, T\}$ (w.l.o.g.) that at the very least "covers" $\mathcal{T}$ in the sense that every $t \in \mathcal{T}$ appears in $\mathcal{C}$. By a corpus we basically mean, here, a giant body of text or a "long" sequence of elements of $\mathcal{T}$. Can we compute the empirical distribution of next tokens for some batch size B? Let's define what we mean clearly: we want to estimate

$$\rho_B(\mathbf{t}, \tau) = \mathbb{P}(\ t_{B+1} = \tau \mid t_1, \dots, t_B\ )$$

empirically via *something* like

$$\hat{\rho}_B(\mathbf{t}, \tau) = \frac{|\{(\mathbf{t}, \tau) \subset \mathcal{C} : |\mathbf{t}| = B\}|}{|\{\mathbf{t} \subset \mathcal{C} : |\mathbf{t}| = B\}|}$$

Why *something* like? The question would be whether *exactly* this empirical value would be *generative* without an underlying structural model. Simply, suppose we have some $\mathbf{t}$, we generate a next τ, but such that $(t_2, \dots, t_B, \tau)$ is not in $\mathcal{C}$. Strictly speaking, we can't then use literal empirical distributions as they would not tell us what comes next, at least as a B sequence. (We might need to cascade downward in sequence size, meaning find the largest $0 < B' = B - h < B$ such that $(t_h, \dots, t_B, \tau)$ can be generative from $\mathcal{C}$; simple occurrence is always defined, so we can do this.)

Clearly $\rho_{B-1}(\mathbf{t}, \tau)$ is a *marginal* relative to $\rho_B(\mathbf{t}, \tau)$, because ρ_B is more specific than ρ_{B-1}:

$$\rho_{B-1}(\mathbf{t}, \tau) = \mathbb{P}(\ t_B = \tau \mid t_1, \dots, t_{B-1}\ ) = \sum_t \rho_0(t) \rho_B\big(t \circ \mathbf{t}, \tau\big)$$

where $\circ$ is a concatenation operator and $\rho_0(t) = \mathbb{P}(T = t)$. Also

$$\hat{\rho}_0(\tau) = \frac{\#\text{ of occurrences of } \tau \text{ in } \mathcal{C}}{|\mathcal{C}|}$$

which we are assured is defined and positive.

So how would we compute $\hat{\rho}_B$? To start, we can do this in a single pass over $\mathcal{C}$ as follows:

(0) Initialize a hashmap $\mathcal{P}_B$ whose keys are B-token strings and values are also hashmaps with single-token keys and whose values are floats (technically `doubles`). We will store $\hat{\rho}_B(\mathbf{t}, \tau) = \mathcal{P}_B[\mathbf{t}][\tau]$. Initialize $c = B$.

(1) Set $\mathbf{t} = \mathcal{C}[c - B : c)$, $\tau = \mathcal{C}[c]$

(2) If $\mathcal{P}_B[\mathbf{t}]$ does not exist, initialize $\mathcal{P}_B[\mathbf{t}]$ as needed, and set $\mathcal{P}_B[\mathbf{t}][\tau] = 1$. Otherwise, if $\mathcal{P}_B[\mathbf{t}][\tau]$ does not exist, set $\mathcal{P}_B[\mathbf{t}][\tau] = 1$. Otherwise increment $\mathcal{P}_B[\mathbf{t}][\tau]$.

(3) Increment c and go back to (1) unless $c = |\mathcal{C}|$, in which case continue to (4).

(4) For all keys $\mathbf{t}$ defined in $\mathcal{P}_B$, compute

$$S(\mathbf{t}) = \sum_{\tau \in \mathcal{P}_B[\mathbf{t}]} \mathcal{P}_B[\mathbf{t}][\tau]$$

and update

$$\mathcal{P}_B[\mathbf{t}][\tau] \leftarrow \mathcal{P}_B[\mathbf{t}][\tau]/S(\mathbf{t})$$

Technically this is a bit like a double-pass algorithm considering the normalization in (4), but it is still $O(|\mathcal{C}|)$. We also might want to reverse the ordering of tokens in the keys of $\mathcal{P}_B$, to enable easier search with some tools that can do bulk return with partial key matching, but that's a detail. This is also embarassingly parallelizable, distributing the right overlapping subsets of $\mathcal{C}$ and merging globally, though we don't outline the details.

Now, we also need to reduce to $P_{B-1}, P_{B-2}, \dots, P_0$ for any hope of prediction. Specifically, we could predict a token τ for which the concatenated subsequence $(\mathbf{t}[2:]) \circ \tau$ does not occur in the corpus. By assumption P_0 exists and is "complete", as the empirical frequencies of tokens in the corpus are defined by virtue of covering. That is, we can always simply sample from the simple occurrence likelihood. Our process could be to take find longest suffix of $\mathbf{t}$ that exists in the corpus,

$$\mathbf{t}[:-k] \quad \text{where} \quad k = \arg \min_{0 \leq k \leq |\mathbf{t}|} \{ |\mathbf{t}[:-k]| : \mathbf{t}[:-k] \subset \mathcal{C} \}$$

and sample from $P_{|\mathbf{t}[:-k]|}[\mathbf{t}[:-k]]$. In the "worst case" $k = |\mathbf{t}|$ and we choose from P_0.

Prediction actually means a $\mathcal{T}$-set of suffix trees may be more suitable, where we access $P_B[\mathbf{t}]$ via following the "reverse" or suffix path

$$t_{|\mathbf{t}|} \to t_{|\mathbf{t}|-1} \circ t_{|\mathbf{t}|} \to \cdots$$

from a (guaranteed-to-exist) root $t_{|\mathbf{t}|}$ to the deepest accessible node. This could also aid in the "marginalization" process. Each node would contain a hashmap representing the distribution over next most likely tokens.

Mountain View CA
Email address: morrowwr@gmail.com