Chart-based noisy channel PLU model

Emily Kellison-Linn

8 Aug 2017

1 Model

The following describes a noisy channel model for computing a distribution over bottom-level PLU sequences given a top-level PLU sequence and a set of model parameters.

1.1 Parameters

Let:

- $a_1, ...a_k \equiv$ the alphabet of k unique PLUs
- $D_1, ...D_k \equiv$ the deletion probabilities for each PLU. D_i corresponds to the probability of top-level PLU a_i being deleted and thus not present in the bottom level at any given index at which a_i appears in s. $D_i \in [0, 1]\ \forall i$.
- $I_1, ...I_k \equiv$ the insertion probabilities for each PLU. I_i corresponds to the probability of top-level PLU a_i being inserted into the bottom level at any given index. $I_i \in [0, 1]\ \forall i$.
- $S_{11}, ...S_{kk} \equiv$ the substitution probabilities for each pair of PLUs. $S_{i,j}$ corresponds to the probability of top-level PLU a_i being replaced by PLU a_j in the bottom level at any given index at which a_i appears.$S_{i,j} \in [0, 1]\ \forall i, j$.
- $P_{11}, ...P_{nk} \equiv$ the prior probabilities of each PLU at each index, based on the properties of the acoustic data.

1.2 Data

Let $s_1, ...s_n \equiv$ the top-level PLU sequence of length n,
where $s_i \in \{a_1, ...a_x\}$ for all $1 \leq i \leq n$.

1.3 Computation

Let $M_{111}, ...M_{n,n,2k+1} \equiv$ the PLU sequence probability chart. **Note: I know we should actually permit M to be of size $M_{n,m,2k+1}$, where m is some number**

slightly larger than n. But under my current understanding of the model I don't understand how that works yet.

Each 1-dimensional vector M_{ij*} of M corresponds to a prefix $s_1, ...s_i - 1$ of s and a prefix of length $j-1$ of the bottom-level PLUs. Each individual cell M_{ijq} corresponds to the probability of the most probable bottom-level PLU sequence of length $j-1$ generated from $s_1, ...s_i - 1$ whose last edit operation is q.

For each M_{ij*}, let the first element correspond to a delete operation (deleting $s_i - 1$); let the next k elements correspond to insert operations (inserting $a_1, ...a_k$ following s_{i-2}); and let the next k elements correspond to substitution operations (substituting $a_1, ...a_k$ for $s_i - 1$). This gives us $|M_{ij*}| = 2k + 1$.

The chart can then be filled iteratively as follows:

$M_{11q} = 1 \; \forall q$ (probabilities for transforming one 0-length string into another)

$$M_{1,j,q} = \begin{cases} max(M_{1,j-1,*}) \times I'_q \times P_{j-1,q'} \;;\; 2 \leq q \leq k+1 \text{ where } q' = q - (k+1) \\ 0 \text{ otherwise} \end{cases}$$

(first row; series of insertions)

$$M_{i,1,q} = \begin{cases} max(M_{i-1,1,*}) \times D_{s_{i-1}} \;;\; q = 1 \\ 0 \text{ otherwise} \end{cases}$$

(first column; series of deletions)

For the general case ($i > 1$ and $j > 1$):

$$M_{i,j,q} = \begin{cases} max(M_{i-1,j,*}) \times D_{s_{i-1}} \;;\; q = 1 & \text{(Delete operation)} \\ max(M_{i,j-1,*}) \times I'_q \times P_{j-1,q''} \;;\; 2 \leq q \leq k+1 \text{ where } q'' = q - 1 & \\ & \text{(Insert operations)} \\ max(M_{i-1,j-1,*}) \times S_{s_{i-1},q'} \times P_{j-1,q'} \;;\; k+2 \leq q \leq 2k+1 \text{ where } q' = q - (k+1) & \\ & \text{(Substitute operations)} \end{cases}$$