Markov Process Lexicon

April 19, 2017

We are interested in modeling the generation of a random lexicon as a Markov process. For the process, we imagine stringing characters together until, at some random point, the string is considered a word. There are thus four qualitatively different states for the Markov process (and degeneracies corresponding to the number of allowable characters). We will need to check

1. the columns sum to 1 (so the transition to the next state is a probability),
2. the largest eigenvalue is equal to 1,
3. the lexicon generated is finite (finite words are generated).

The possible states are the top node (before adding the first letter), adding a character that does not finish a word, adding a character that finishes a word, and the bottom node (corresponding to a finished word that has no suffixes). For N characters, the transition matrix P for this process has a shape of $2(N+1)\times 2(N+1)$ (2 edge nodes, N characters that don't finish words, N characters that could finish words). If we imagine each character as equally likely and completable words being completed with probability α, then the transition matrix can be expressed as

$$P_{ij} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1/N & 1/2N & \cdots & (1-\alpha)/2N & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 1/N & 1/2N & \cdots & (1-\alpha)/2N & \cdots & 0 \\ 0 & 1/2N & \cdots & (1-\alpha)/2N & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 1/2N & \cdots & (1-\alpha)/2N & \cdots & 0 \\ 0 & 0 & \cdots & \alpha & \cdots & 1 \end{bmatrix}. \tag{1}$$

We've got to unpack this a little. The first row represents the top node state - no state should be able to return to the top node, so it is filled with zeros. The next N rows correspond to the N characters that don't end words. The top node leads only to these to

exclude words that are only one character long, but it can lead to any of the characters with equal probability, hence the $1/N$ populating that segment of the first column. The next N rows are the N characters that end words. Characters that do and do not end words may transition to each other (e.g., ∘end̄ might transition to ∘end̄i on its way to ∘end̄ing̅s̄). The latter selects the next character with equal probability ($1/2N$), but the former must retain the possibility of actually ending the branch (as no word can be constructed from ∘end̄ing̅s̄•). We denote the probability of ending a branch as α. This α factor is removed uniformly from the transition probabilities away from a character that end words, hence $(1-\alpha)/2N$. Finally, the last row corresponds to the bottom node state, that cannot transition to anywhere.

The columns can similarly be partitioned into those four categories, where we define the following vectors

$$\mathbf{v_T} = \begin{bmatrix} 0 \\ 1/N \\ \vdots \\ 1/N \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{v_{NW}} = \begin{bmatrix} 0 \\ 1/2N \\ \vdots \\ 1/2N \\ 1/2N \\ \vdots \\ 1/2N \\ 0 \end{bmatrix} \quad \mathbf{v_W} = \begin{bmatrix} 0 \\ (1-\alpha)/2N \\ \vdots \\ (1-\alpha)/2N \\ (1-\alpha)/2N \\ \vdots \\ (1-\alpha)/2N \\ \alpha \end{bmatrix} \quad \mathbf{v_B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{2}$$

Clearly there are many linear dependencies between the columns (as there are N each of $\mathbf{v_{NW}}$ and $\mathbf{v_W}$). An additional linear dependence that is of note is

$$\mathbf{v_W} = \alpha \mathbf{v_B} + (1-\alpha)\mathbf{v_{NW}} \tag{3}$$

We proceed now to evaluate how each of these vectors is acted upon by P.

$$P_{ij}\mathbf{v_T} = \begin{bmatrix} 0 \\ N(1/N)(1/2N) \\ \vdots \\ N(1/N)(1/2N) \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/2N \\ \vdots \\ 1/2N \\ 0 \end{bmatrix} = \mathbf{v_{NW}} \tag{4}$$

$$P_{ij}\mathbf{v_{NW}} = \begin{bmatrix} 0 \\ N(1/2N)^2 + N(1/2N)((1-\alpha)/2N) \\ \vdots \\ N(1/2N)^2 + N(1/2N)((1-\alpha)/2N) \\ N(1/2N)\alpha \end{bmatrix} = \begin{bmatrix} 0 \\ (2-\alpha)/4N \\ \vdots \\ (2-\alpha)/4N \\ \alpha/2 \end{bmatrix} \tag{5}$$

$$= \frac{2-\alpha}{2}\mathbf{v_{NW}} + \frac{\alpha}{2}\mathbf{v_B} \tag{6}$$

$$P_{ij}\mathbf{v_W} = \begin{bmatrix} 0 \\ N(1/2N)((1-\alpha)/2N) + N((1-\alpha)/2N)^2 \\ \vdots \\ N(1/2N)((1-\alpha)/2N) + N((1-\alpha)/2N)^2 \\ N\alpha(1-\alpha)/2N + \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ (1-\alpha)(2-\alpha)/4N \\ \vdots \\ (1-\alpha)(2-\alpha)/4N \\ \alpha(3-\alpha)/2 \end{bmatrix} \tag{7}$$

$$= \frac{2-\alpha}{2}\mathbf{v_W} + \frac{\alpha}{2}\mathbf{v_B} \tag{8}$$

$$P_{ij}\mathbf{v_B} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} = \mathbf{v_B} \tag{9}$$

Note that $\mathbf{v_T}$ transitions into the "not word" state, which is what we wanted. Next, observe that the bottom node state transitions only to itself, and is therefore our first eigenvector, and has an eigenvalue of one.

$$\mathbf{v_B} \cdot \mathbf{v_B} = 1 \tag{10}$$
$$\mathbf{v_T} \cdot \mathbf{v_T} = 1/N \tag{11}$$
$$\mathbf{v_{NW}} \cdot \mathbf{v_{NW}} = 1/2N \tag{12}$$
$$\mathbf{v_W} \cdot \mathbf{v_W} = (1 - 2\alpha + (1+2N)\alpha^2)/2N \tag{13}$$
$$\mathbf{v_B} \cdot \mathbf{v_T} = 0 \tag{14}$$
$$\mathbf{v_B} \cdot \mathbf{v_{NW}} = 0 \tag{15}$$
$$\mathbf{v_B} \cdot \mathbf{v_W} = \alpha \tag{16}$$
$$\mathbf{v_T} \cdot \mathbf{v_{NW}} = 1/2N \tag{17}$$
$$\mathbf{v_T} \cdot \mathbf{v_W} = (1-\alpha)/2N \tag{18}$$
$$\mathbf{v_{NW}} \cdot \mathbf{v_W} = (1-\alpha)/2N \tag{19}$$
$$\tag{20}$$

$$\mathbf{u_1} = \mathbf{v_B} \tag{21}$$

$$\mathbf{u_2} = \mathbf{v_W} - \frac{\mathbf{v_B} \cdot \mathbf{v_W}}{\mathbf{v_B} \cdot \mathbf{v_B}}\mathbf{v_B} = \mathbf{v_W} - \alpha\mathbf{v_B} \tag{22}$$

$$P_{ij}\mathbf{u_2} = P\left(\mathbf{v_W} - \alpha\mathbf{v_B}\right) = P\mathbf{v_W} - \alpha P\mathbf{v_B} \tag{23}$$

Substituting from eqs 43 and 44

$$= \frac{2-\alpha}{2}\mathbf{v_W} + \frac{\alpha}{2}\mathbf{v_B} - \alpha\mathbf{v_B} \tag{24}$$

$$= \frac{2-\alpha}{2}\mathbf{v_W} - \frac{\alpha}{2}\mathbf{v_B} \tag{25}$$

Try

$$\mathbf{v_2} = \mathbf{v_B} - \mathbf{v_{NW}} = \begin{bmatrix} 0 \\ -1/2N \\ \vdots \\ -1/2N \\ -1/2N \\ \vdots \\ -1/2N \\ 1 \end{bmatrix} \tag{26}$$

$$P_{ij}\mathbf{v_2} = \begin{bmatrix} 0 \\ N(-1/2N)(1/2N) + N(-1/2N)((1-\alpha)/2N) \\ \vdots \\ N(-1/2N)(1/2N) + N(-1/2N)((1-\alpha)/2N) \\ N(-1/2N)\alpha + 1 \end{bmatrix} \tag{27}$$

$$= \begin{bmatrix} 0 \\ ((2-\alpha)/2)(-1/2N) \\ \vdots \\ ((2-\alpha)/2)(-1/2N) \\ (2-\alpha)/2 \end{bmatrix} = \frac{2-\alpha}{2}\begin{bmatrix} 0 \\ -1/2N \\ \vdots \\ -1/2N \\ 1 \end{bmatrix} = \frac{2-\alpha}{2}\mathbf{v_2} \tag{28}$$

$$\mathbf{v_2} \cdot \mathbf{v_2} = 2N\frac{1}{4N^2} + 1 = \frac{1+2N}{2N} \tag{29}$$

Alternatively,

$$P\mathbf{v_2} = P(\mathbf{v_B} - \mathbf{v_{NW}}) = \mathbf{v_B} - (\frac{2-\alpha}{2}\mathbf{v_{NW}} + \frac{\alpha}{2}\mathbf{v_B}) \tag{30}$$

$$= \frac{2-\alpha}{2}(\mathbf{v_B} - \mathbf{v_{NW}}) = \frac{2-\alpha}{2}\mathbf{v_2} \tag{31}$$

Now we may write our starting state $\mathbf{v_{NW}}$ as

$$\mathbf{v_{NW}} = \mathbf{v_B} - \mathbf{v_2} \tag{32}$$

Now, since

$$P\mathbf{v_{NW}} = \mathbf{v_B} - \frac{2-\alpha}{2}\mathbf{v_2}, \tag{33}$$

repeated application of the transition matrix yields

$$P^n\mathbf{v_{NW}} = \mathbf{v_B} - \left(\frac{2-\alpha}{2}\right)^n \mathbf{v_2}. \tag{34}$$

As long as $0 < \alpha \leq 1$, we know that $(2-\alpha)/2 < 1$, implying that

$$\lim_{n\to\infty} P^n\mathbf{v_{NW}} = \mathbf{v_B}, \tag{35}$$

which means that all our branches must eventually end and so we may expect finite words.

A Errata

We are interested in modeling the generation of a random lexicon as a Markov process. For the process, we imagine stringing characters together until, at some random point, the string is considered a word. There are thus four qualitatively different states for the Markov process (and degeneracies corresponding to the number of allowable characters). We will need to check

1. the columns sum to 1 (so the transition to the next state is a probability),
2. the largest eigenvalue is equal to 1,
3. the lexicon generated is finite (finite words are generated).

The possible states are the top node (before adding the first letter), adding a character that does not finish a word, adding a character that finishes a word, and the bottom node (corresponding to a finished word that has no suffixes). For N characters, the transition matrix P for this process has a shape of $2(N+1)\times 2(N+1)$ (2 edge nodes, N characters that don't finish words, N characters that could finish words). If we imagine each character as equally likely and completable words being completed with probability α, then the transition matrix can be expressed as

$$P_{ij} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1/N & 1/2N & \cdots & (1-\alpha)/2N & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 1/N & 1/2N & \cdots & (1-\alpha)/2N & \cdots & 0 \\ 0 & 1/2N & \cdots & (1-\alpha)/2N & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 1/2N & \cdots & (1-\alpha)/2N & \cdots & 0 \\ 0 & 0 & \cdots & \alpha & \cdots & 1 \end{bmatrix}. \tag{36}$$

We've got to unpack this a little. The first row represents the top node state - no state should be able to return to the top node, so it is filled with zeros. The next N rows correspond to the N characters that don't end words. The top node leads only to these to exclude words that are only one character long, but it can lead to any of the characters with equal probability, hence the $1/N$ populating that segment of the first column. The next N rows are the N characters that end words. Characters that do and do not end words may transition to each other (e.g., $\circ\text{end}\bar{}$ might transition to $\circ\text{end}\bar{}\text{i}$ on its way to $\circ\text{end}\bar{}\text{ing}\bar{}\bar{\text{s}}$). The latter selects the next character with equal probability ($1/2N$), but the former must retain the possibility of actually ending the branch (as no word can be constructed from $\circ\text{end}\bar{}\text{ing}\bar{}\bar{\text{s}}\bullet$). We denote the probability of ending a branch as α. This α factor is removed uniformly from the transition probabilities away from a character that end words, hence $(1-\alpha)/2N$. Finally, the last row corresponds to the bottom node state, that cannot transition to anywhere.

The columns can similarly be partitioned into those four categories, where we define the following vectors

$$\mathbf{v_T} = \begin{bmatrix} 0 \\ 1/N \\ \vdots \\ 1/N \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{v_{NW}} = \begin{bmatrix} 0 \\ 1/2N \\ \vdots \\ 1/2N \\ 1/2N \\ \vdots \\ 1/2N \\ 0 \end{bmatrix} \quad \mathbf{v_W} = \begin{bmatrix} 0 \\ (1-\alpha)/2N \\ \vdots \\ (1-\alpha)/2N \\ (1-\alpha)/2N \\ \vdots \\ (1-\alpha)/2N \\ \alpha \end{bmatrix} \quad \mathbf{v_B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{37}$$

Clearly there are many linear dependencies between the columns (as there are N each of $\mathbf{v_{NW}}$ and $\mathbf{v_W}$). An additional linear dependence that is of note is

$$\mathbf{v_W} = \alpha\mathbf{v_B} + (1-\alpha)\mathbf{v_{NW}} \tag{38}$$

We proceed now to evaluate how each of these vectors is acted upon by P.

$$P_{ij}\mathbf{v_T} = \begin{bmatrix} 0 \\ N(1/N)(1/2N) \\ \vdots \\ N(1/N)(1/2N) \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/2N \\ \vdots \\ 1/2N \\ 0 \end{bmatrix} = \mathbf{v_{NW}} \tag{39}$$

$$P_{ij}\mathbf{v_{NW}} = \begin{bmatrix} 0 \\ N(1/2N)^2 + N(1/2N)((1-\alpha)/2N) \\ \vdots \\ N(1/2N)^2 + N(1/2N)((1-\alpha)/2N) \\ N(1/2N)\alpha \end{bmatrix} = \begin{bmatrix} 0 \\ (2-\alpha)/4N \\ \vdots \\ (2-\alpha)/4N \\ \alpha/2 \end{bmatrix} \tag{40}$$

$$= \frac{2-\alpha}{2}\mathbf{v_{NW}} + \frac{\alpha}{2}\mathbf{v_B} \tag{41}$$

$$P_{ij}\mathbf{v_W} = \begin{bmatrix} 0 \\ N(1/2N)((1-\alpha)/2N) + N((1-\alpha)/2N)^2 \\ \vdots \\ N(1/2N)((1-\alpha)/2N) + N((1-\alpha)/2N)^2 \\ N\alpha(1-\alpha)/2N + \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ (1-\alpha)(2-\alpha)/4N \\ \vdots \\ (1-\alpha)(2-\alpha)/4N \\ \alpha(3-\alpha)/2 \end{bmatrix} \tag{42}$$

$$= \frac{2-\alpha}{2}\mathbf{v_W} + \frac{\alpha}{2}\mathbf{v_B} \tag{43}$$

$$P_{ij}\mathbf{v_B} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} = \mathbf{v_B} \tag{44}$$

Note that $\mathbf{v_T}$ transitions into the "not word" state, which is what we wanted. Next, observe that the bottom node state transitions only to itself, and is therefore our first eigenvector, and has an eigenvalue of one.

$$\mathbf{v_B} \cdot \mathbf{v_B} = 1 \tag{45}$$

$$\mathbf{v_T} \cdot \mathbf{v_T} = 1/N \tag{46}$$

$$\mathbf{v_{NW}} \cdot \mathbf{v_{NW}} = 1/2N \tag{47}$$

$$\mathbf{v_W} \cdot \mathbf{v_W} = (1 - 2\alpha + (1+2N)\alpha^2)/2N \tag{48}$$

$$\mathbf{v_B} \cdot \mathbf{v_T} = 0 \tag{49}$$

$$\mathbf{v_B} \cdot \mathbf{v_{NW}} = 0 \tag{50}$$

$$\mathbf{v_B} \cdot \mathbf{v_W} = \alpha \tag{51}$$

$$\mathbf{v_T} \cdot \mathbf{v_{NW}} = 1/2N \tag{52}$$

$$\mathbf{v_T} \cdot \mathbf{v_W} = (1-\alpha)/2N \tag{53}$$

$$\mathbf{v_{NW}} \cdot \mathbf{v_W} = (1-\alpha)/2N \tag{54}$$

$$\tag{55}$$

$$\mathbf{u_1} = \mathbf{v_B} \tag{56}$$

$$\mathbf{u_2} = \mathbf{v_W} - \frac{\mathbf{v_B} \cdot \mathbf{v_W}}{\mathbf{v_B} \cdot \mathbf{v_B}} \mathbf{v_B} = \mathbf{v_W} - \alpha \mathbf{v_B} \tag{57}$$

$$P_{ij}\mathbf{u_2} = P\left(\mathbf{v_W} - \alpha \mathbf{v_B}\right) = P\mathbf{v_W} - \alpha P \mathbf{v_B} \tag{58}$$

Substituting from eqs 43 and 44

$$= \frac{2-\alpha}{2}\mathbf{v_W} + \frac{\alpha}{2}\mathbf{v_B} - \alpha \mathbf{v_B} \tag{59}$$

$$= \frac{2-\alpha}{2}\mathbf{v_W} - \frac{\alpha}{2}\mathbf{v_B} \tag{60}$$

Try

$$\mathbf{v_2} = \mathbf{v_B} - \mathbf{v_{NW}} = \begin{bmatrix} 0 \\ -1/2N \\ \vdots \\ -1/2N \\ -1/2N \\ \vdots \\ -1/2N \\ 1 \end{bmatrix} \tag{61}$$

$$P_{ij}\mathbf{v_2} = \begin{bmatrix} 0 \\ N(-1/2N)(1/2N) + N(-1/2N)((1-\alpha)/2N) \\ \vdots \\ N(-1/2N)(1/2N) + N(-1/2N)((1-\alpha)/2N) \\ N(-1/2N)\alpha + 1 \end{bmatrix} \tag{62}$$

$$= \begin{bmatrix} 0 \\ ((2-\alpha)/2)(-1/2N) \\ \vdots \\ ((2-\alpha)/2)(-1/2N) \\ (2-\alpha)/2 \end{bmatrix} = \frac{2-\alpha}{2}\begin{bmatrix} 0 \\ -1/2N \\ \vdots \\ -1/2N \\ 1 \end{bmatrix} = \frac{2-\alpha}{2}\mathbf{v_2} \tag{63}$$

$$\mathbf{v_2} \cdot \mathbf{v_2} = 2N\frac{1}{4N^2} + 1 = \frac{1+2N}{2N} \tag{64}$$

Alternatively,

$$P\mathbf{v_2} = P(\mathbf{v_B} - \mathbf{v_{NW}}) = \mathbf{v_B} - (\frac{2-\alpha}{2}\mathbf{v_{NW}} + \frac{\alpha}{2}\mathbf{v_B}) \tag{65}$$

$$= \frac{2-\alpha}{2}(\mathbf{v_B} - \mathbf{v_{NW}}) = \frac{2-\alpha}{2}\mathbf{v_2} \tag{66}$$