Morphology Segmentation

For each word $w_i \in W$, j split points are sampled from a Geometric distribution $G(\theta)$ where θ = 0.5.

For each split point, the word w_i is split into two parts at the split point p_j such that $w_i = x_i + y_i$.

Here, $x_i \in M_x$ is the stem belonging to the stem morpheme set and $y_i \in M_y$ is the suffix belonging to suffix morpheme set for the word w_i.

We draw two distributions over all possible stems and suffix from a Dirichlet process with concentration parameter α and base distribution G_x and G_y respectively.

$$X \mid \alpha, G_x \sim DP(\alpha, G_x)$$
$$Y \mid \alpha, G_y \sim DP(\alpha, G_y)$$

where,

G_x and G_y are geometric distributions over the length of the string. The pdf of a geometric distribution is given as follows:

$$P(x|\theta) = \theta\,(1-\theta)^{x-1}$$

where,

x = number of failures before success (in our case, length of the string).

We integrate out θ by using the conjugate prior Beta Distribution with parameters α *and* β

$$p(\theta \mid \alpha, \beta) = \frac{1}{B(\alpha,\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}$$

Posterior Geometric under Beta Prior

$$p(\theta \mid X, \alpha, \beta) = B(\alpha + n, \beta + \sum_i (x_i - 1))$$

where, X = all the data points
n = total count of data points

Posterior Predictive

$$p(x \mid \alpha', \beta') = \frac{B(\alpha' + 1, \beta' + x - 1)}{B(\alpha', \beta')}$$

Now, we have,

$$X \mid \alpha, BG_x \sim DP(\alpha, BG_x)$$
$$Y \mid \alpha, BG_y \sim DP(\alpha, BG_y)$$

where, $BG_x \; and \; BG_y$ are the posterior predictive (Beta-Geometric Distribution)

Dirichlet Process as Chinese Restaurant Process

We treat words as customers and the morphemes as tables in the restaurant. Therefore, we consider two restaurants, one for stem and other for suffix i.e. two Dirichlet Processes.

The Dirichlet Process with concentration parameter α and base distribution G is defined as

$$p(w_i, z_i | W_{-i}, Z_{-i}, \alpha, \alpha', \beta') = CRP(w_i, z_i, \alpha)\, p(w_i | \alpha'_{z_i}, \beta'_{z_i})$$

where,

$$CRP(w_i, z_i, \alpha) = \frac{n_{z_{-i}}}{N + \alpha - 1}, \; if \; z_i \; is \; an \; existing \; table$$
$$\frac{\alpha}{N + \alpha - 1}, \; if \; z_i \; is \; a \; new \; table$$

where,

z is the latent variable representing a table k

$n_{z_{-i}}$ is the count of data in a table k

N is the total data

$$p(w_i|\alpha'_{z_i}, \beta'_{z_i}) = \frac{B(\alpha'_{z_i}+1, \beta'_{z_i}+l-1)}{B(\alpha'_{z_i}, \beta'_{z_i})}$$

where,

l is the length of w_i

$\alpha'_{z_i} = \alpha + n$

$\beta'_{z_i} = \beta + \sum_{-i}(w_{-i} - 1)$

Inference

To determine the split point of a new word, we generate all possible splits for the given word and pick the split pair with the maximum probability.

$$argmax\ p(w_{new} = x_{new} + y_{new}|\ X, Y)$$

The probability measure is given by:

$$p(w_{new} = x_{new} + y_{new}|\ X, Y) = p(x_{new}|\ X)\ p(y_{new}\ |\ Y)$$

where,

$$p(x_{new}|\ X, \alpha) = \frac{n_{z_{new}}}{N+\alpha},\ if\ z_{new} \in M_x$$
$$\frac{\alpha\ .\ p(x_{new}\ |\ \alpha', \beta')}{N+\alpha},\ else$$

where, $p(x_{new}|\ \alpha',\ \beta') = \frac{B(\alpha'+1, \beta'+l-1)}{B(\alpha', \beta')}$

Similarly,

$$p(y_{new}|\ Y\ , \alpha) = \frac{n_{z_{new}}}{N+\alpha},\ if\ z_{new} \in M_y$$
$$\frac{\alpha\ .\ p(y_{new}\ |\ \alpha', \beta')}{N+\alpha},\ else$$

where, $p(y_{new}|\ \alpha',\ \beta') = \frac{B(\alpha'+1, \beta'+l-1)}{B(\alpha', \beta')}$