A Obtaining the weight update equations in the case of cosine similarity

To obtain the weight update equations for the input and output vectors of our model in each iteration of stochastic gradient descent, we must find the gradient of the error function at a given training example, which may be considered a document, n-gram pair.

Let:

$$E = -\log\sigma\left(\alpha\cos\theta_{w_o}\right) - \sum_{w_n\in W_{neg}}\log\sigma\left(-\alpha\cos\theta_{w_n}\right) \quad (1)$$

where:

$$\cos\theta_w = \frac{\boldsymbol{v}_d^T\boldsymbol{v}_w}{\|\boldsymbol{v}_d\|\|\boldsymbol{v}_w\|} \quad (2)$$

be the objective function at a single training example (d, w_o). Then, to find the gradient of E first differentiate E with respect to $\cos\theta_w$:

$$\frac{\partial E}{\partial\cos\theta_w} = \alpha\left(\sigma\left(\alpha\cos\theta_w\right) - t\right) \quad (3)$$

where $t = 1$ if $w = w_o$; 0 otherwise. We then obtain the derivative of E w.r.t. the output n-gram vectors:

$$\frac{\partial E}{\partial\boldsymbol{v}_w} = \frac{\partial E}{\partial\cos\theta_w}\cdot\frac{\partial\cos\theta_w}{\partial\boldsymbol{v}_w} \quad (4)$$

$$\frac{\partial E}{\partial\boldsymbol{v}_w} = \alpha\left(\sigma\left(\alpha\cos\theta_w\right) - t\right)\cdot\left(\frac{\boldsymbol{v}_d}{\|\boldsymbol{v}_d\|\|\boldsymbol{v}_w\|} - \frac{\boldsymbol{v}_w\left(\boldsymbol{v}_d^T\boldsymbol{v}_w\right)}{\|\boldsymbol{v}_d\|\|\boldsymbol{v}_w\|^3}\right) \quad (5)$$

This leads to the following weight update equation for the output vectors:

$$\boldsymbol{v}_w^{(new)} = \boldsymbol{v}_w^{(old)} - \eta\frac{\partial E}{\partial\boldsymbol{v}_w} \quad (6)$$

where η is the learning rate. This equation needs to be applied to all $w\in\{w_o\}\cup W_{neg}$ in each iteration.

Next, the errors are backpropagated and the input document vectors are updated. Differentiating E with respect to $\boldsymbol{v}_d$:

$$\frac{\partial E}{\partial\boldsymbol{v}_d} = \sum_{w\in\{w_o\}\cup W_{neg}}\frac{\partial E}{\partial\cos\theta_w}\cdot\frac{\partial\cos\theta_w}{\partial\boldsymbol{v}_d} \quad (7)$$

$$= \sum_{w\in\{w_o\}\cup W_{neg}}\alpha\left(\sigma\left(\alpha\cos\theta_w\right) - t\right)\cdot\left(\frac{\boldsymbol{v}_w}{\|\boldsymbol{v}_d\|\|\boldsymbol{v}_w\|} - \frac{\boldsymbol{v}_d\left(\boldsymbol{v}_d^T\boldsymbol{v}_w\right)}{\|\boldsymbol{v}_d\|^3\|\boldsymbol{v}_w\|}\right) \quad (8)$$

Thus, we obtain the weight update equation for the input vector in each iteration:

$$\boldsymbol{v}_d^{(new)} = \boldsymbol{v}_d^{(old)} - \eta\frac{\partial E}{\partial\boldsymbol{v}_d} \quad (9)$$

B Weight update equations in the case of dot product

This section contains the weight update equations for the input and output vectors of the dot product model in each iteration of stochastic gradient descent.

The following weight update equations for the output vectors:

$$\boldsymbol{v}_w^{(new)} = \boldsymbol{v}_w^{(old)} - \eta\left(\sigma\left(\boldsymbol{v}_d^T\boldsymbol{v}_w\right) - t\right)\cdot\boldsymbol{v}_d \quad (10)$$

where $t = 1$ if $w = w_o$; 0 otherwise, needs to be applied to all $w\in\{w_o\}\cup W_{neg}$ in each iteration.

The following weight update equation needs to be applied to the input vector in each iteration:

$$\boldsymbol{v}_d^{(new)} = \boldsymbol{v}_d^{(old)} - \eta\sum_{w\in\{w_o\}\cup W_{neg}}\left(\sigma\left(\boldsymbol{v}_d^T\boldsymbol{v}_w\right) - t\right)\cdot\boldsymbol{v}_w \quad (11)$$

C Weight update equations in the case of L2R dot product

This section contains the weight update equations for the input and output vectors of the L2R dot product model in each iteration of stochastic gradient descent.

The following weight update equations for the output vectors:

$$\boldsymbol{v}_w^{(new)} = \boldsymbol{v}_w^{(old)} - \eta\left(\sigma\left(\boldsymbol{v}_d^T\boldsymbol{v}_w\right) - t\right)\cdot\boldsymbol{v}_d - \eta\lambda\boldsymbol{v}_w \quad (12)$$

where $t = 1$ if $w = w_o$; 0 otherwise, needs to be applied to all $w\in\{w_o\}\cup W_{neg}$ in each iteration.

The following weight update equation needs to be applied to the input vector in each iteration:

$$\boldsymbol{v}_d^{(new)} = \boldsymbol{v}_d^{(old)} - \eta\sum_{w\in\{w_o\}\cup W_{neg}}\left(\sigma\left(\boldsymbol{v}_d^T\boldsymbol{v}_w\right) - t\right)\cdot\boldsymbol{v}_w - \eta\lambda\boldsymbol{v}_d \quad (13)$$