A LSTM Layer

The LSTM layer is defined as follows:

$$
\begin{aligned}
i_t &= \sigma(W_{ii}e_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi}), \\
f_t &= \sigma(W_{if}e_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}), \\
g_t &= \tanh(W_{ig}e_t + b_{ig} + W_{hc}h_{(t-1)} + b_{hg}), \\
o_t &= \sigma(W_{io}e_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho}), \\
c_t &= f_t * c_{(t-1)} + i_t * g_t, \\
h_t &= o_t * \tanh(c_t),
\end{aligned}
$$

where e_t is the input word embeddings.

B Matrix-based Equations

Given the matrix representations of two adjacent sentences (e.g. H_i and H_j), the similarity matrix between LSTM states of these two sentences is defined as follows:

$$SIM = |H_i \cdot H_j^T|,$$

where SIM is the similarity matrix and H_j^T is the transpose of matrix H_j.

$$I_{index}, J_{index} = \texttt{ARGMAX}\left(SIM\right),$$

$$\vec{u}, \vec{v} = H_i[I_{index}], H_j[J_{index}],$$

where the ARGMAX function determines the indices (the row index I_{index} and the column index J_{index}) of the maximum element of the SIM matrix. These indices point out to word vectors of H_i and H_j that are considered for representing the sentence relation.

C Convolution Layer

A convolution operation involves applying filter w (i.e. a vector of weight parameters) to the vector of similarities of k continuity degrees among adjacent sentences in order to encode local transitions of the salient topic:

$$\vec{c} = \tanh(w^T \cdot L_{t:t+k-1} + b_t),$$

where $L_{t:t+k-1}$ denotes the k elements in the vector representation of degrees of continuity and b_t is the bias. Notice that we use a wide convolution, as opposed to narrow, to ensure that the filters reach entire elements of an input vector, including the boundaries. We do this by performing zero-padding, where elements located out of boundaries are assumed to be zero.

D QWK

To calculate QWK, between two sets of scores, a weight matrix W is constructed as follows:

$$W_{ij} = \frac{(i-j)^2}{(N-1)^2},$$

where i is the rating assigned by a human annotator and j is the rating assigned by a system. N is the number of possible ratings. A matrix O is calculated such that $O_{i,j}$ is the number of essays that receive a rating i by the human annotator and a rating j by the system. The last matrix is E that is calculated by the outer product of the histogram vectors of the human and system ratings. The matrix E is then normalized such that the sum of the elements in E and the sum of the elements in O are the same. QWK is calculated using the matrices W, O, and E as follows:

$$QWK = 1 - \frac{\sum_{i,j} W_{ij}O_{ij}}{\sum_{ij} W_{ij}E_{ij}}.$$