Answers to the Question 3

Here $\lambda_{j+1}^{(t)}$ is the loss at word $j+1$ of the target. We need to find the derivative $\frac{\partial \lambda_{j+1}^{(t)}}{\partial h_n^{(s)}}$ where $h_n^{(s)}$ is the hidden state at word n of source. So using the chain rule we get:

$$\frac{\partial \lambda_{j+1}^{(t)}}{\partial h_n^{(s)}} = \frac{\partial \lambda_{j+1}^{(t)}}{\partial \lambda_j^{(t)}} \frac{\partial \lambda_j^{(t)}}{\partial h_{j-1}^{(t)}} \cdots \frac{\partial \lambda h_{n+1}^{(s)}}{\partial \lambda h_n^{(s)}}$$

The above equation represents that we need to apply chain rule from the loss of current target word to the loss of the $n+1$ source word.

To find the shortest backpropagation path we need to consider the length of the sequences. Specifically, we need to consider j^{th} term of the target sequence and $m^{(s)} - n$ term of the source sequence (assuming $m^{(s)}$ is the length of source sequence. So total terms on the shortest backpropagation: $m^{(s)} - n + j$

Answer to the Question 4

For the attention model, the derivative

$$\frac{\partial \lambda_{j+1}^{(t)}}{\partial z_n} = \frac{\partial \lambda_{j+1}^{(t)}}{\partial c_j} \frac{\partial c_j}{\partial z_n}$$

Here c_j is taken from the textbook (Einstein's book equation 18.39),

where $c_j = \sum_{n=1}^{M^{(s)}} \alpha_{j \to n} z_n$

So, $$\frac{\partial \lambda_{j+1}^{(t)}}{\partial z_n} = \frac{\partial \lambda_{j+1}^{(t)}}{\partial c_j} \frac{\partial \sum_{n=1}^{M^{(s)}} \alpha_{j \to n'} z_{n'}}{\partial z_n}$$

As we are differentiating the second term with respect to z_n, in the shortest path we don't need to consider $n' = n$. So, the shortest path $= \frac{\partial \lambda_{j+1}^{(t)}}{\partial c_j} \frac{\partial \sum_{n' \neq n}^{M^{(s)}} \alpha_{j \to n'} z_{n'}}{\partial z_n}$