CS 224N

2.d) ii.

If we did not use masking, the loss would increase, since we would be adding $M-T$ terms of the form $\hat{y}^{(t)}$, $t>T$.

Also, the gradient would change. As we saw before, if $\hat{y} = \text{softmax}(\theta)$ and $J = CE$ loss, $\frac{\partial J}{\partial \theta} = (\hat{y} - y)$. So not only y has a 1 in the "NULL" position, but $\hat{y}$ is filled with nonzero elements as well. This first error δ backpropagates and will end up affecting most parameters gradients.

Masking solves the problem by having an $m^{(t)} = \mathbb{1}\{t \leq T\}$ multiplying each term of the loss (and, thus, the gradients), for each pos $t = 1, \ldots, M$. That would set to 0 both loss and gradients for $t = T+1, \ldots, M$ for $(x^{(t)}, y^{(t)})$, solving the issue (i.e., the new gradients and loss would be equivalent as if we didn't do the augmentation).