Encoder-Decoder: Training Objective

For output word y_i we have:

- estimated conditional distribution $\hat{p}_i = \frac{\exp t_i}{\sum \exp t_j}$ (softmax function)
- unknown true distribution p_i, we lay $p_i \equiv \mathbf{1}\left[y_i\right]$

Cross entropy $\approx$ distance of $\hat{p}$ and p:

$$\mathcal{L} = H(\hat{p}, p) = \mathbf{E}_p\left(-\log \hat{p}\right)$$