Page 8,

To understand how to switch our training from log-likelihood objective to RL scenario, lets lets look at both from the mathematical point of view. Log-likelihood estimation means maximizing the sum $\sum_{i=1}^{N} \log p_{model}(y_i|x_i)$ by tweaking models parameter, which is exactly the same as minimization of KL-divergence between the data probability distribution and and probability distribution parameterized by the model, which could be written as maximisation of $\mathbb{E}_{x\sim p_{data}} \log p_{model}(x)$

On the other hand, the REINFORCE method from chapter 9 has the objective to maximize $\mathbb{E}_{s\sim data, a\sim\pi(a|s)} Q(s,a) \log \pi(a|s)$

Later on the same page

6. Estimate of the gradient $\nabla J = \sum_T Q \nabla \log p(T)$

Page 9

Switching to argmax mode makes the decoder process fully deterministic and provides the baseline for REINFORCE policy gradient in the formula

$$\nabla J = \mathbb{E}[(Q(s) - b(s)) \nabla \log p(a|s)]$$