A PYTHIA training parameters

Both encoder and decoder of PYTHIA consist of 2-layers with 512 hidden units. During training, we use dropout with probability 0.2 and scheduled sampling with probability 0.5 (Bengio et al., 2015). All models were trained on an 8-core machine with an NVIDIA 1080 Ti graphics processing unit (GPU). The batch size was 32 and the network weights were optimised using Adam (Kingma and Ba, 2015) with a learning rate of 10^{-3} and gradient clipping of 5.

B LM training parameters

The language modelling LSTM network consists of 2-layers with 1024 hidden units and an equally sized character embedding space. The parameters were trained using Adam with a learning rate of $2 \cdot 10^{-3}$, a decay of 0.95, gradient norm clipping of 5, and dropout probability 0.2 for the inputs and the hidden layers.