Results

Training the switching probabilities using a seq to seq model

Changing the context that is shared across the recurrent units

Perplexity	Data	RNNLM	DLM
No pre-training	Dev	94.1	93.65
	Test	78.8	75.49

Other variants

Tried changing the loss function

Results

Perplexity	Data	RNNLM	DLM
No pre-training	Dev	94.1	93.65
	Test	78.8	75.49

Other variants

Training the switching probabilities using a seq to seq model

Changing the context that is shared across the recurrent units

Tried changing the loss function

Byte-level RNNLM[1]

Idea: Use a more lower level of representation of input.
i.e. use bytes instead of characters as input

 [1] Gillick . et al. Multilingual Processing from Bytes. 2015