

LM hyper-parameter settings for 100k & 90k Europarl corpora	
Parameter	Value
Word embedding size	256
Hidden layer size	512
Number of layers	2
Dropout probability	0.5
Sampled softmax samples	25
L2 regularization beta	1e-5
Gradient norm upper bound	10.0
Optimizer	ADAM
Learning rate	0.001
Learning rate annealing steps	2
Learning rate annealing factor	0.9
Unrolled time-steps	35
Batch size	32
Warm-up epochs	6
Stagnant epochs before early stopping	20
Max. training epochs	50

Table 1: LM hyper-parameter choices during corpus construction phase.