Sentence similarity classifier hyper-parameter settings	
Parameter	**Value**
Word embedding size	256
Hidden layer size	50
Number of layers	2
Number of data buckets	8
Dropout probability	0.5
L2 regularization beta	1e-5
Gradient norm upper bound	12.0
Optimizer	ADAM
Learning rate	0.0001
Learning rate annealing steps	2
Learning rate annealing factor	0.9
Batch size	64
Warm-up epochs	3
Stagnant epochs before early stopping	10
Max. training epochs	50

Table 1: Sentence similarity classifier hyper-parameter choices during pre-training on human-annotated and synthetic data.