A Results on GloVe.840B embeddings

			WordSim										Analogy		
method	hyper-parameters		MEN	MC	MTurk	RW	R&G	SCWS	Slex	WSR	WSS	Macro	GL	MSYN	Micro
GloVe.840B	-	-	.805	.788	.693	.462	.769	.632	.408	.688	.803	.636	82.3	80.7	81.9
SUM-F	$F = 0.5$M	-	.768	**.829**	**.746**	**.566**	**.817**	**.668**	.402	.578	**.806**	.653	53.3	**71.2**	58.0
SUM-H	-	$H = 0.5$M	**.803**	.801	.710	.485	.753	.630	.433	**.687**	**.806**	.643	59.1	63.2	60.2
KVQ-H	-	$H = 0.5$M	.773	.798	.691	.452	.718	.605	.377	.559	.756	.608	**67.0**	69.2	**67.6**
SUM-FH	$F = 1.0$M	$H = 0.5$M	**.803**	.760	.715	.518	.728	.646	**.438**	.671	.788	**.674**	41.7	49.6	43.8
KVQ-FH	$F = 1.0$M	$H = 0.5$M	.781	.787	.704	.479	.760	.613	.406	.568	.724	.647	58.8	61.3	59.5
SUM-F	$F = 0.2$M	-	.731	**.809**	**.710**	**.526**	**.777**	**.642**	.369	.482	.762	.617	40.3	**66.2**	47.2
SUM-H	-	$H = 0.2$M	.750	.733	.660	.458	.655	.604	.370	.587	.754	.599	38.1	56.4	42.9
KVQ-H	-	$H = 0.2$M	.702	.619	.603	.444	.681	.553	.327	.451	.646	.556	**43.0**	59.5	**47.4**
SUM-FH	$F = 1.0$M	$H = 0.2$M	**.761**	.722	.701	.505	.657	.625	**.385**	**.628**	**.773**	**.640**	25.5	40.1	29.3
KVQ-FH	$F = 1.0$M	$H = 0.2$M	.737	.698	.638	.452	.690	.577	.347	.559	.697	.600	35.7	50.3	39.6

Table 9: Results of model shrinkage experiments. The reconstruction target is GloVe.840B.

Table 9 shows the results of model shrinkage experiments whose reconstruction target is GloVe.840B.

B Training settings for downstream tasks

model				type	crf_tagger
				dropout	0.5
	text_field-_embedder	token-_embedders	tokens	embedding_dim	300
				trainable	true
			tokens_characters	embedding_dim	16
			encoder	type	cnn
				embedding_dim	16
				num_filters	128
				ngram_filter_sizes	[3]
				conv_layer_activation	relu
	encoder			type	lstm
				input_size	428
				hidden_size	200
				num_layers	2
				dropout	0.5
				bidirectional	true
iterator				batch_size	64
trainer	optimizer			type	adam
				num_epochs	75
				grad_norm	5.0
				patience	25
dataset_reader	token_indexers	tokens		lowercase_tokens	false

Table 10: Training settings for the AllenNLP's implementation for the NER experiments.

model				type	esim
				dropout	0.5
	text_field_embedder	token_embedders	tokens	embedding_dim	300
				trainable	true
	encoder			type	lstm
				input_size	300
				hidden_size	300
				num_layers	1
				bidirectional	true
	similarity_function			type	dot_product
	projection_feedforward			input_dim	2400
				hidden_dims	300
				num_layers	1
				activations	relu
	inference_encoder			type	lstm
				input_size	300
				hidden_size	300
				num_layers	1
				bidirectional	true
	output_feedforward			input_dim	2400
				num_layers	1
				hidden_dims	300
				activations	relu
				dropout	0.5
	output_logit			input_dim	300
				num_layers	1
				hidden_dims	3
				activations	linear
iterator				batch_size	32
trainer	optimizer			type	adam
				lr	0.0004
				num_epochs	75
				grad_norm	10.0
				patience	5
dataset_reader	token_indexers	tokens		lowercase_tokens	true

Table 11: Training settings for the AllenNLP's implementation for the TE experiments.