A Appendix

Dataset		Batch	Lr	Layer	Dropout	CLS DIM	Weight decay
SST (Socher et al., 2013)		25	1e-4	4	0.3	300	1e-4
MTL-16 † (Liu et al., 2017)	Apparel Baby Books Camera DVD Electronics Health IMDB Kitchen Magazines MR Music Software Sports Toys Video	20	1e-3	4	0.3	300	1e-4
PTB POS (Marcus et al., 1993)		64	5e-4	4	0.2	300	1e-4
CoNLL03 (Sang and Meulder, 2003)		64	5e-4	4	0.5	300	5e-5
OntoNotes NER (Pradhan et al., 2012)		64	3e-4	4	0.3	300	1e-4
SNLI (Bowman et al., 2015)		64	1e-4	2	0.3	600	5e-5

Table 7: Additional information of hyper-parameters, "Batch" means the batch size, "Lr" means the learning rate, "Emb" means fine-tune the embedding or fix it, "Dropout" is the dropout ratio, "CLS DIM" means the hidden size of classifier, "Weight decay" is the L2 regularization penalty on parameters excluding the embedding.

Transformer Baseline The common implementation of Transformer (for Machine Translation) is not suitable for tasks experimented in this work, to obtain a meaningful baseline, we need to do several modifications.

The first point is incorporating with the pre-trained word embedding like GloVe, we need to use the learnable position embedding and remove the coefficient $\sqrt{\text{H DIM}}$ before the embedding.

The second point is the regularization, besides the dropout in MultiAtt and FFN, we also add the dropout in the embedding layer and classifier which is widely used in LSTM models. The weight decay is also important for avoiding overfitting, we add the weight decay on all parameters except the embedding layer.

The third point is the warm-up of the optimizer, it is not designed for small tasks, an alternative solution is to initialize the weights in a small range likes $N(0.0, 0.05)$.