Supplemental Material for “Dynamic Compositionality in Recursive Neural Networks with Structure-aware Tag Representations”

[†]Taeuk Kim, [†]Jihun Choi, [‡]Daniel Edmiston, [†]Sanghwan Bae, [†]Sang-goo Lee
[†]Department of Computer Science and Engineering, Seoul National University, Seoul, Korea
{taeuk, jhchoi, sanghwan, sglee}@europa.snu.ac.kr
[‡]Department of Linguistics, University of Chicago, Chicago, IL, USA
danedmiston@uchicago.edu

Clustered Tag Set

In this section, we provide a specification of our proposed clustered tag set.

First, word-level tags are grouped together according to the universal POS tagset (Petrov, Das, and McDonald 2012), which was originally proposed to integrate tag sets for different languages. To be specific, we utilize the pre-defined tag set for PennTreeBank-style POS tags. For some tags which are not specified in the pre-defined tag set but exist in real parse trees, we manually add the rules such that the tags can be clustered with syntactically similar tags. As a result, all word-level tags are grouped into 12 categories. The details of the grouping information are presented in table 1.

Second, we define our own grouping strategy for the phrase-level tags as a pre-defined one does not exist in case of the phrase tags. This is similar to that for the word-level tags. The newly specified 11 phrase-level tag groups are presented in table 2. Note that we distinguish word-level and phrase-level tag sets so that each tag set can reflect the characteristics of the leaf-nodes and non-leaf nodes of parse trees respectively.

Detailed Experimental Settings

Dataset Statistics

In the quantitative analysis portion of our paper, we have tested our model on 6 sentence-level datasets in total. Here, we report the simple statistics of the datasets. For details, refer to table 3. Note that we use the subtrees of parse trees in addition to the whole parse trees when training models for SST-2 and SST-5, following the standard in the literature. We have parsed sentences in the datasets by utilizing the Stanford PCFG parser (Klein and Manning 2003), even in cases where the datasets provide parse trees. If the original parse trees from the datasets are not matched to our newly parsed trees, we exclude the subtrees of the original ones from training.

Task-specific Model Settings and Variations

We present the detailed settings and the selected hyperparameter values of our models used in the experiments. The specification includes various model configurations such as the size of tag embeddings ($\mathbf{d}_{\mathrm{T}}$), the size of the hidden and cell state of our word tree-LSTM ($\mathbf{d}_h$), and the dimension

Original Tags	Groups
NN, NP, NNP, NNS, NNPS, NX, WHNP	N
PRP, PRP$, WP, WP$, PRP, PRP$	PN
VP, VB, VBD, VBG, VBN, VBP, VBZ, MD	V
ADJP, JJ, JJR, JJS	ADJ
ADVP, WHADVP, RB, RBR, RBS, WRB	ADV
DT, EX, PDT, WDT	DET
IN	ADP
CC	CONJ
CD	NUM
!, #, $, ?, comma(,), colon(:), period(.), quotation marks(“, ”), -LRB-, -RRB-, LST, PRN	PUNC
PRT, PP, TO, POS, RP	PRT
Other word-level tags	X

Table 1: The (modified) universal POS tagset.

Original Tags	Groups
NP, @NP, NX, @NX, WHNP, @WHNP	NP
VP, @VP	VP
ADJP, @ADJP, WHADJP, @WHADJP	ADJP
ADVP, @ADVP, WHADVP, @WHADVP	ADVP
S, @S, SBAR, @SBAR, SQ, @SQ SINV, @SINV	S
ROOT	ROOT
CONJP, @CONJP	CONJP
QP, @QP	NUMP
LST, @LST	PUNCP
PRT, @PRT, PP, @PP, WHPP, @WHPP	PRTP
Other phrase-level tags	XP

Table 2: The phrase-level tag grouping.

Dataset	$\mathbf{d_c}$	$\boldsymbol{l}$	# Train	# Dev	# Test
SST-2	2	19	93,517	872	1,821
SST-5	5	18	300,192	1,101	2,210
MR	2	20	10,662	-	CV
SUBJ	2	23	10,000	-	CV
TREC	6	10	5,452	-	500
SNLI	3	11	550,152	10,000	10,000

Table 3: Summary statistics for the sentence classification datasets. $\mathbf{d}_c$: Number of classes. $\boldsymbol{l}$: Average sentence length. **# Train**, **# Dev**, **# Test**: The number of training, validation, and test data respectively. **CV**: 10-fold cross validation.

Dataset	$\mathbf{d}_{\mathrm{T}}$	$\mathbf{d}_h$	$\mathbf{d}_s$	**Optimizer**	**L**	**B**	**E**	**W**	**D**	**Word embedding fine-tuning**
SST-2	50	100	500	Adadelta	1	64	20	1e-5	0.5	Y
SST-5	100	300	1200	Adadelta	1	64	10	5e-6	0.5	Y
MR	25	100	500	Adadelta	1	64	30	1e-5	0.5	Y
SUBJ	25	150	1000	Adadelta	1	64	40	2e-5	0.5	Y
TREC	50	300	800	Adadelta	1	64	40	5e-5	0.5	Y
SNLI	100	300	1200	Adam	1e-3	64	10	1e-5	0.1	N

Table 4: The task-specific settings of our best models for each dataset. $\mathbf{d}_{\mathrm{T}}$: The dimension size of tag embeddings. $\mathbf{d}_h$: The dimension size of the hidden and cell state of the word tree-LSTM. $\mathbf{d}_s$: The dimension size of each task-specific classifier. **L**: Learning rate. **B**: Batch size. **E**: Max epoch size. **W**: Weight decay rate. **D**: Dropout (drop) probability. **word embedding fine-tuning**: Whether word embeddings are fine-tuned during training or not (Y: fine-tuned, N: fixed).

size of the task-specific classifier ($\mathbf{d}_s$). The details are reported in table 4.

In addition to the above task-specific settings, we introduce some model variations which we experimented with. In the experiment on SST-5, we applied layer normalization (Ba, Kiros, and Hinton 2016) to the input values of the gate functions in the word-level tree-LSTM. Moreover, the 'ROOT' tag group is merged into the 'S' tag group in case of SST-5, MR, and SUBJ. For SNLI, we used the original syntactic tags instead of our clustered tag set. All the variations help in boosting the model performance for respective cases.

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Klein, D., and Manning, C. D. 2003. Accurate unlexicalized parsing. In *ACL*, 423–430.

Petrov, S.; Das, D.; and McDonald, R. 2012. A universal part-of-speech tagset. In *LREC*.