Supplement Material for Higher-order Syntactic Attention Network for Longer Sentence Compression

Hidetaka Kamigaito♡, Katsuhiko Hayashi♢, Tsutomu Hirao♠ and Masaaki Nagata♠

♡ Institute of Innovative Research, Tokyo Institute of Technology
♢ Institute of Scientific and Industrial Research, Osaka University
♠ NTT Communication Science Laboratories, NTT Corporation
kamigaito@lr.pi.titech.ac.jp, katsuhiko-h@sanken.osaka-u.ac.jp,
{hirao.tsutomu,nagata.masaaki}@lab.ntt.co.jp

		ALL					LONG				
		$\mathbf{F_1}$	ROUGE			ΔC	$\mathbf{F_1}$	ROUGE			ΔC
			1	*2*	*L*			*1*	*2*	*L*	
Tagger		79.8	79.7	70.3	79.5	-1.5	76.4	75.5	65.8	75.1	-2.6
Tagger+ILP		76.9	76.8	66.0	76.5	-2.7	75.4	72.3	60.3	71.7	-2.9
Bi-LSTM		78.6	79.4	70.4	79.1	-0.4	74.8	75.8	66.3	75.3	-1.0
Bi-LSTM-Dep		78.9	80.0	71.1	79.7	$\mathbf{-0.1}$	74.5	76.2	66.9	75.7	+0.6
Attn		79.1	79.2	70.3	79.0	-1.1	75.5	76.0	66.6	75.6	-1.4
Base		79.7	79.2	70.5	78.9	-1.8	76.1	76.0	67.0	75.5	-2.0
HiSAN-Dep	($\mathbf{d} = \{1\}$)	79.3	79.9	70.9	79.6	-0.7	75.5	76.4	67.0	76.0	-1.1
	($\mathbf{d} = \{1,2\}$)	79.7	80.6	71.7	80.3	-0.5	76.0	77.1	67.8	76.7	-0.9
HiSAN-Dep	($\mathbf{d} = \{1,2,3\}$)	79.8	81.0	72.0	80.7	$\mathbf{-0.1}$	76.0	77.7	68.3	77.2	-0.5
	($\mathbf{d} = \{1,2,4\}$)	79.7	81.3	72.5	81.0	+0.3	75.9	77.9	68.7	77.4	$\mathbf{-0.2}$
	($\mathbf{d} = \{1,2,3,4\}$)	79.7	80.6	71.8	80.3	-0.3	75.7	77.2	68.0	76.8	-0.6
	($\mathbf{d} = \{1\}$)	80.4	81.0	71.9	80.7	-0.6	77.2	77.8	68.1	77.3	-1.2
	($\mathbf{d} = \{1,2\}$)	80.6	81.4	72.2	81.1	-0.5	77.3	78.1	68.3	77.6	-1.3
HiSAN	($\mathbf{d} = \{1,2,3\}$)	**80.7**	**82.2**	**73.1**	**81.9**	+0.2	**77.6**	79.3	**69.8**	**78.9**	-0.4
	($\mathbf{d} = \{1,2,4\}$)	80.5	82.0	72.9	81.7	$\mathbf{+0.1}$	77.4	**79.4**	**69.8**	**78.9**	$\mathbf{-0.2}$
	($\mathbf{d} = \{1,2,3,4\}$)	80.6	81.2	72.1	81.0	-0.8	77.4	78.1	68.4	77.7	-1.2

Table 1: Results of automatic evaluation on the small training data set (8,000 sentences)[1]. **ALL** and **LONG**, respectively represent the results in all sentences and long sentences (longer than average length 28) in the test dataset. $\mathbf{d}$ represents the groups of d-length dependency chains. Bold results indicate the best scores. All results are reported as the average scores of five trials.

	$\mathbf{F_1}$	ROUGE			ΔC	AVG
		1	*2*	*L*		
HiSAN-Dep						
$\mathbf{d} = \{1\}$	80.9	80.6	72.0	80.4	-1.5	62.5
$\mathbf{d} = \{1,2\}$	81.1	81.1	72.3	80.8	-1.3	62.8
$\mathbf{d} = \{1,2,3\}$	81.3	82.0	73.3	81.7	-0.8	63.5
$\mathbf{d} = \{1,2,4\}$	81.4	82.5	74.0	82.2	-0.3	**64.0**
$\mathbf{d} = \{1,2,3,4\}$	81.2	81.8	73.2	81.5	-0.8	63.4
HiSAN						
$\mathbf{d} = \{1\}$	81.3	81.5	72.6	81.2	-1.3	63.1
$\mathbf{d} = \{1,2\}$	81.6	81.9	73.1	81.6	-1.2	63.4
$\mathbf{d} = \{1,2,3\}$	81.6	82.6	73.8	82.3	-0.5	**64.0**
$\mathbf{d} = \{1,2,4\}$	81.5	82.5	73.7	82.2	-0.5	63.9
$\mathbf{d} = \{1,2,3,4\}$	81.8	81.8	73.1	81.5	-1.5	63.3

Table 2: Results in development dataset on the small training dataset (8,000 sentences). **AVG** represents the average of all metrics. All results are reported as the average scores of five trials.

[1] In the small setting, the dropout rate was set to 0.65.

	$\mathbf{F_1}$	ROUGE			ΔC	AVG
		1	*2*	*L*		
HiSAN-Dep						
$\mathbf{d} = \{1\}$	84.0	82.4	75.9	82.2	-2.9	64.3
$\mathbf{d} = \{1, 2\}$	84.4	82.7	76.2	82.5	-3.0	64.5
$\mathbf{d} = \{1, 2, 3\}$	84.2	82.4	75.8	82.2	-3.1	64.3
$\mathbf{d} = \{1, 2, 4\}$	84.4	82.8	76.4	82.7	-2.9	**64.7**
$\mathbf{d} = \{1, 2, 3, 4\}$	83.7	82.4	75.7	82.1	-2.6	64.3
HiSAN						
$\mathbf{d} = \{1\}$	84.2	82.4	76.0	82.2	-3.2	64.3
$\mathbf{d} = \{1, 2\}$	84.2	82.8	76.3	82.6	-2.7	64.6
$\mathbf{d} = \{1, 2, 3\}$	84.1	82.7	76.1	82.4	-2.8	64.5
$\mathbf{d} = \{1, 2, 4\}$	84.2	83.3	76.8	83.1	-2.3	**65.0**
$\mathbf{d} = \{1, 2, 3, 4\}$	84.3	82.7	76.2	82.5	-2.8	64.6

Table 3: Results in development dataset on the large training dataset (200,000 sentences). **AVG** represents the average of all metrics. All results are reported as the average scores of five trials.

		ALL		LONG		DEPTH	
		$\mathbf{F_1}$	ΔC	$\mathbf{F_1}$	ΔC	$\mathbf{F_1}$	ΔC
Tagger		79.7	-0.8	76.2	-2.4	79.0	-1.7
Tagger+ILP		77.4	-2.7	73.7	-3.2	76.1	-4.2
Bi-LSTM		78.8	$\mathbf{-0.1}$	75.1	-0.9	78.5	-0.6
Bi-LSTM-Dep		79.0	$\mathbf{+0.1}$	74.8	-0.5	78.1	$\mathbf{-0.3}$
Attn		79.3	-0.9	75.7	-1.3	79.1	-1.1
Base		79.8	-1.5	76.4	-1.9	79.2	-1.7
HiSAN-Dep	($\mathbf{d} = \{1\}$)	79.5	-0.4	75.8	-0.9	79.0	-0.8
HiSAN-Dep	($\mathbf{d} = \{1, 2\}$)	80.0	-0.2	76.4	-0.7	79.5	-0.6
	($\mathbf{d} = \{1, 2, 3\}$)	80.1	$\mathbf{+0.1}$	76.5	-0.3	79.6	+0.4
	($\mathbf{d} = \{1, 2, 4\}$)	80.7	+0.6	76.2	$\mathbf{+0.0}$	79.1	$\mathbf{+0.3}$
	($\mathbf{d} = \{1, 2, 3, 4\}$)	**80.9**	$\mathbf{-0.1}$	76.0	-1.0	79.1	-0.5
HiSAN	($\mathbf{d} = \{1\}$)	80.5	-0.2	77.4	-1.0	80.0	-0.4
	($\mathbf{d} = \{1, 2\}$)	80.8	$\mathbf{-0.1}$	77.6	-1.1	80.3	-0.6
	($\mathbf{d} = \{1, 2, 3\}$)	**80.9**	+0.6	**77.8**	-0.1	80.4	+0.4
	($\mathbf{d} = \{1, 2, 4\}$)	80.7	+0.4	**77.8**	$\mathbf{-0.0}$	80.4	$\mathbf{+0.3}$
	($\mathbf{d} = \{1, 2, 3, 4\}$)	**80.9**	-0.5	77.7	-1.0	**80.5**	-0.5

Table 4: Macro-average of the automatic evaluation results on the small training data set (8,000 sentences). **ALL**, **LONG** and **DEPTH**, respectively represent the results in all sentences, long sentences (longer than average length 28) and sentences with deep dependency trees (deeper than average depth 8) in the test dataset. $\mathbf{d}$ represents the groups of d-length dependency chains. Bold results indicate the best scores. The compression ratio of all gold sentences, longer gold sentences and deeper gold sentences are 43.7, 32.7 and 36.8, respectively. All results are reported as the average scores of five trials.

		ALL		LONG		DEPTH	
		$\mathbf{F_1}$	$\mathbf{\Delta C}$	$\mathbf{F_1}$	$\mathbf{\Delta C}$	$\mathbf{F_1}$	$\mathbf{\Delta C}$
Tagger		83.0	-3.0	80.6	-2.8	83.1	-3.1
Tagger+ILP		79.6	-4.5	76.2	-4.0	78.5	-5.2
Bi-LSTM		82.2	-2.2	79.3	-2.1	81.7	-2.1
Bi-LSTM-Dep		82.6	-2.2	80.1	-1.9	82.0	-2.0
Attn		82.9	-2.4	80.2	-2.2	82.3	-2.1
Base		83.1	-2.4	80.6	-2.3	82.6	-2.4
HiSAN-Dep	($\mathbf{d} = \{1\}$)	83.2	-2.3	80.5	-1.9	82.7	-2.4
HiSAN-Dep	($\mathbf{d} = \{1,2\}$)	83.1	-2.4	81.0	-2.2	82.5	-2.6
	($\mathbf{d} = \{1,2,3\}$)	83.4	-2.5	80.8	-2.3	83.0	-2.5
	($\mathbf{d} = \{1,2,4\}$)	83.3	-2.4	80.9	-2.1	82.5	-2.7
	($\mathbf{d} = \{1,2,3,4\}$)	83.1	-2.2	80.5	-2.1	82.5	-2.5
HiSAN	($\mathbf{d} = \{1\}$)	83.4	-2.9	81.2	-2.6	**83.3**	-2.7
	($\mathbf{d} = \{1,2\}$)	**83.5**	-2.2	**81.3**	-2.0	83.0	-2.1
	($\mathbf{d} = \{1,2,3\}$)	83.3	-2.2	81.1	-2.1	83.0	-2.1
	($\mathbf{d} = \{1,2,4\}$)	**83.5**	$\mathbf{-1.7}$	81.2	$\mathbf{-1.7}$	**83.3**	$\mathbf{-1.8}$
	($\mathbf{d} = \{1,2,3,4\}$)	83.1	-2.3	81.0	-2.3	82.8	-2.3

Table 5: Macro-average of the automatic evaluation results on the large training data set (200,000 sentences). $\mathbf{ALL}$, $\mathbf{LONG}$ and $\mathbf{DEPTH}$, respectively represent the results in all sentences, long sentences (longer than average length 28) and sentences with deep dependency trees (deeper than average depth 8) in the test dataset. $\mathbf{d}$ represents the groups of d-length dependency chains. Bold results indicate the best scores. The compression ratio of all gold sentences, longer gold sentences and deeper gold sentences are 43.7, 32.7 and 36.8, respectively. All results are reported as the average scores of five trials.