Multi-Granularity Self-Attention for Neural Machine Translation

Yang Wei 🐷 51184506043

godweiyang@gmail.com

godweiyang.com

Computer Science and Technology

East China Normal University

Multi-Granularity Self-Attention for Neural Machine Translation

Jie Hao*
Florida State University
haoj8711@gmail.com

Xing Wang
Tencent AI Lab
brightxwang@tencent.com

Shuming Shi
Tencent AI Lab
shumingshi@tencent.com

Jinfeng Zhang
Florida State University
jinfeng@stat.fsu.edu

Zhaopeng Tu
Tencent AI Lab
zptu@tencent.com

Motivations

- SANs generally focus on disperse words and ignore continuous phrase patterns, which have proven essential in both SMT and NMT.
- The power of multiple heads in SANs is not fully exploited.
- Thus this paper (MG-SA) assigns several attention heads to attend over phrase fragments at each granularity.

Framework

- word-level $\rightarrow$ phrase-level memory:

$$H_g = F_h(H).$$

- single head self-attention:

$$\begin{array}{c} Q^h, K^h, V^h = HW_Q^h, H_g W_K^h, H_g W_V^h \\ O^h = \text{ATT}(Q^h, K^h)V^h \end{array}.$$

- final output of MG-SA:

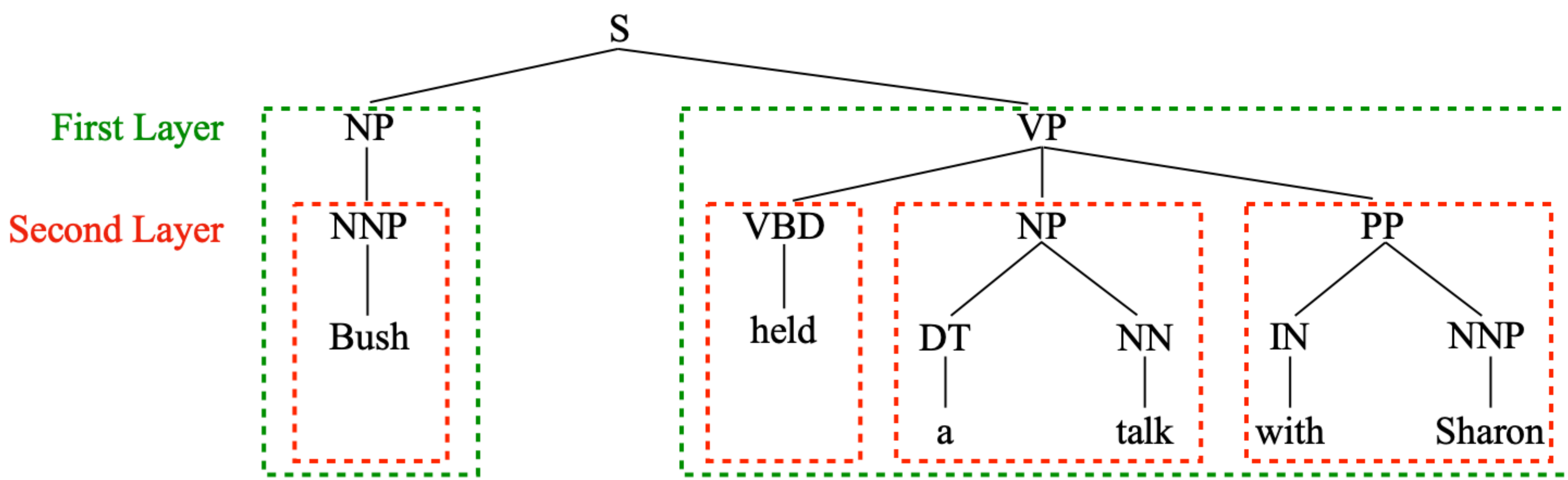
$$\text{MG-SA}(H) = [O^1, \ldots, O^N].$$

Phrase Partition

- split sentence x into M phrases:

$$P_x = (p_1, \ldots .p_M).$$

- partition strategies: **n-gram** or **syntactic**.



Phrase Composition

- phrase representation:

$$g_m = \mathrm{COM}(p_m),$$

where $\mathrm{COM}(\cdot)$ is the composition function with shared parameters to all phrases (e.g. CNN, LSTM and SAN).

- phrase-level memory:

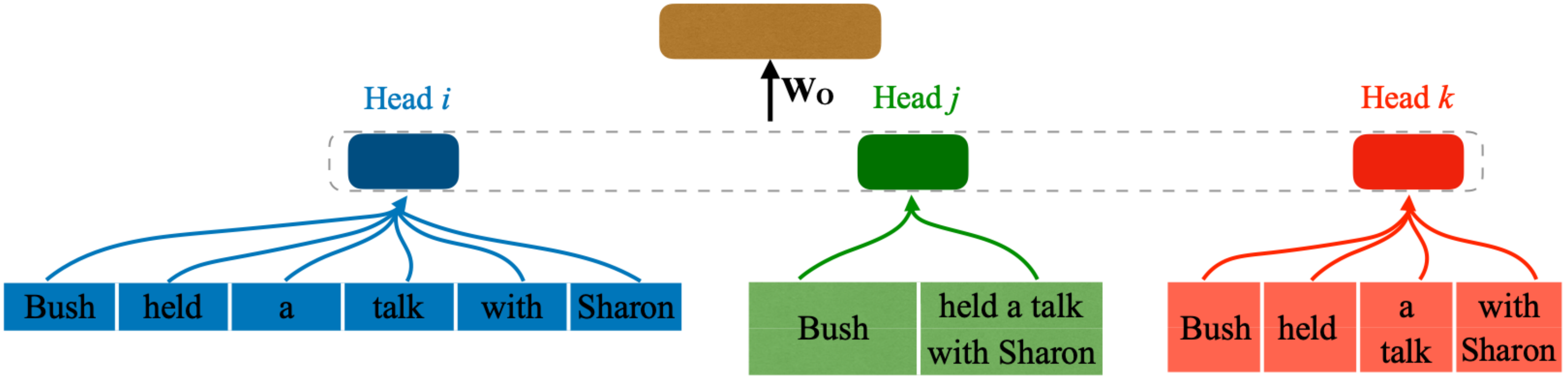
$$G_x = (g_1, \ldots, g_M).$$

Phrase Interaction

- model the latent structure of the phrase sequence:

$$H_g = \mathrm{REC}(G_x),$$

where $\mathrm{REC}(\cdot)$ is the recurrence function (e.g. LSTM and ON-LSTM).



Training

- phrase tag supervision:

$$p = softmax(W_t g_i + b_t)$$
$$\mathcal{L}_{tag} = -\sum_{i=1}^{M} t_i \log p(t_i) \quad .$$

- training loss:

$$\mathcal{L} = -\sum_{i=1}^{L} y_i \log P(y_i) + \lambda \mathcal{L}_{tag}.$$

Three Questions

- Does the integration of the proposed MG-SA into the state-of-the-art TRANSFORMER improve the translation quality in terms of the BLEU score?
- Does the proposed MG-SA promote the generation of the target phrases?
- Does MG-SA capture more phrase information at the various granularity levels?

Phrase Composition

Phrase Modeling	**# Para.**	**Speed**	**BLEU**
n/a	88.0M	1.28	27.31
MAX-POOLING	88.0M	1.27	27.56
SANS	90.4M	1.26	**27.69**
LSTM	96.1M	1.14	27.58

Encoder Layers

Encoder Layers	# Para.	Speed	BLEU
$[1-6]$	90.4M	1.26	27.69
$[1-3]$	89.2M	1.27	27.74
$[1]$	88.4M	1.28	**27.83**

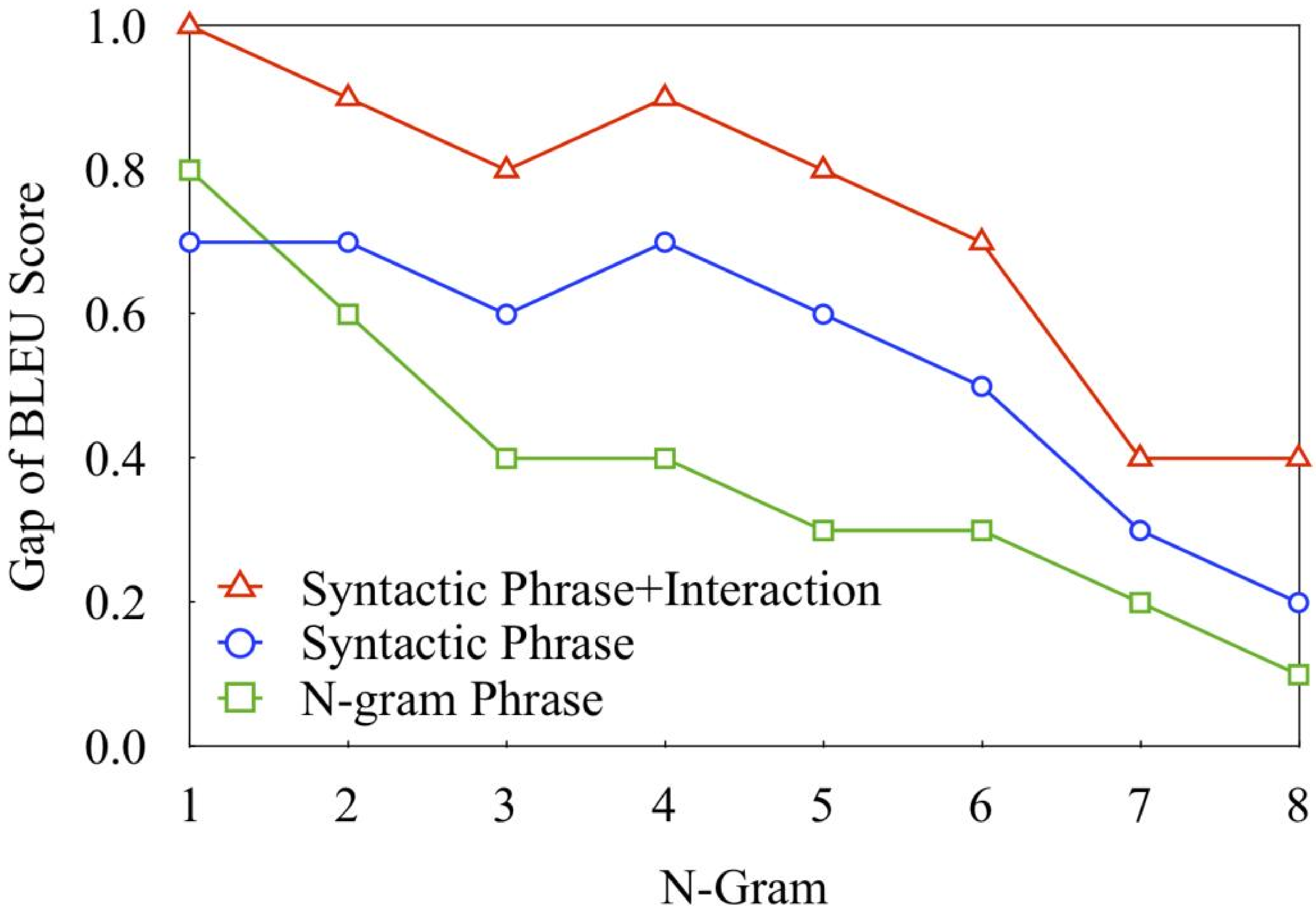
Phrase Partition, Tag Supervision and Phrase Interaction

#	Model Architecture	# Para.	Speed	BLEU	Δ
1	Transformer-Base	88.0M	1.28	27.31	-
2	+ N-gram Phrase	88.4M	1.28	27.83	+0.52
3	+ Syntactic Phrase	88.4M	1.24	28.01	+0.70
4	+ Syntactic Phrase + $\mathcal{L}_{tag}$	88.4M	1.23	28.07	+0.76
5	+ LSTM Interaction	89.5M	1.20	28.14	+0.83
6	+ ON-LSTM Interaction	89.9M	1.19	**28.28**	+0.97

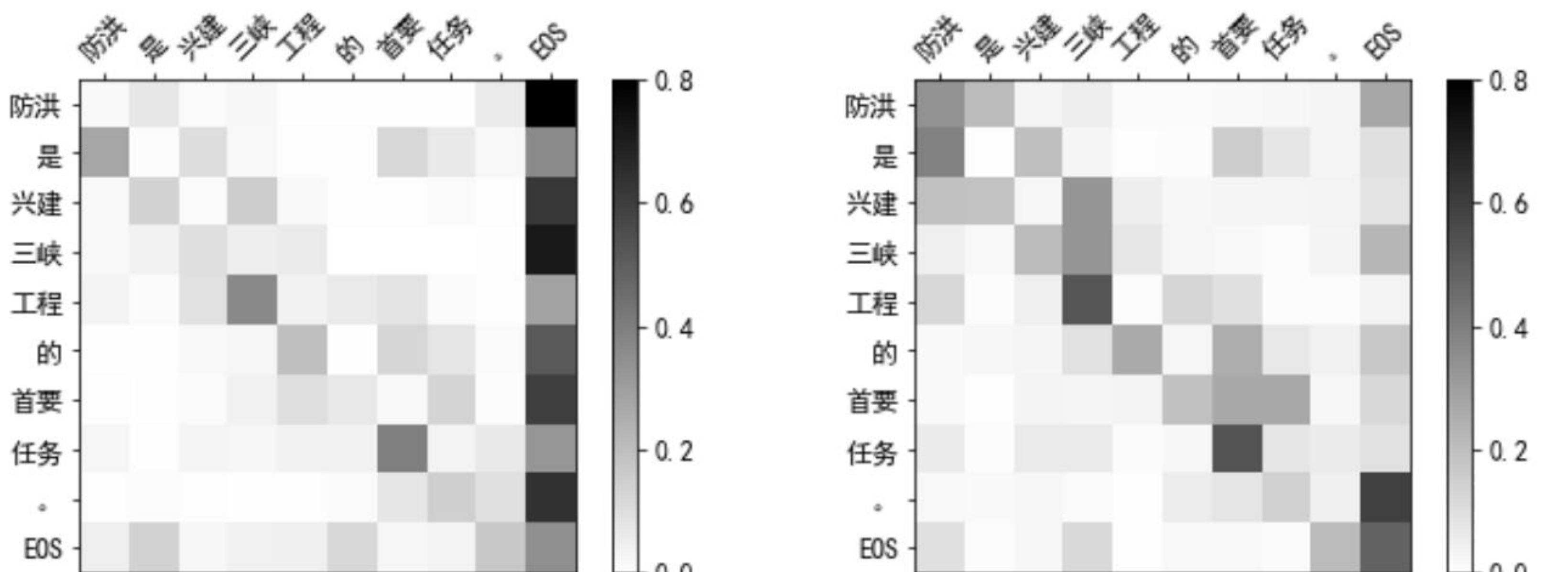
Main Results

Architecture	En⇒De		Zh⇒En					
	# Para.	BLEU	# Para.	MT03	MT04	MT05	MT06	Avg
Existing NMT systems								
Vaswani et al. (2017)	213M	28.4	n/a	n/a	n/a	n/a	n/a	n/a
Zhang et al. (2019)	n/a	n/a	n/a	40.45	42.76	40.09	39.67	40.74
Our NMT systems								
TRANSFORMER-BASE	88.0M	27.31	73.4M	41.88	44.48	42.21	41.93	42.60
+MG-SA	89.9M	28.28$^{\Uparrow}$	75.3M	43.98$^{\Uparrow}$	45.60$^{\Uparrow}$	44.28$^{\Uparrow}$	44.00$^{\Uparrow}$	44.46
TRANSFORMER-BIG	264.1M	28.58	234.8M	45.30	46.49	45.21	44.87	45.47
+MG-SA	271.5M	29.01$^{\uparrow}$	242.2M	45.76$^{\uparrow}$	46.81$^{\uparrow}$	45.77$^{\uparrow}$	46.48$^{\Uparrow}$	46.21

Phrasal Pattern Evaluation



Visualization of Attention



(a) Vanilla Multi-Head Self-Attention

(b) Multi-Granularity Self-Attention

Multi-Granularity Phrases Evaluation

#	Model	Label Granularity: Large → Small					
		Voice	Tense	TSS	SPC	POS	Avg
	Pre-Trained NMT Encoder						
1	BASE	73.38	73.73	72.72	92.81	93.73	81.27
2	N-Gram Phrase	73.06	72.83	72.11	96.42	96.34	82.15
3	Syntactic Phrase	73.37	73.62	75.60	96.72	96.68	83.19
4	Syntactic Phrase + Interaction	73.20	74.78	75.24	96.78	96.56	83.31
	Train From Scratch						
5	BASE	83.46	85.39	83.44	96.35	96.12	88.95
6	N-Gram Phrase	83.55	85.62	85.21	96.23	96.17	89.36
7	Syntactic Phrase	84.70	87.52	97.42	96.95	96.24	92.57
8	Syntactic Phrase + Interaction	86.45	87.65	99.07	96.99	96.40	93.31

Multi-Granularity Phrases Evaluation

- Models trained from scratch consistently outperform NMT encoder probing on all tasks.
- The models with syntactic information significantly perform better than those models without incorporating syntactic information.
- For NMT probing, the proposed models outperform the baseline model especially on relative small granularity of phrases information.
- Models trained from scratch achieve more improvements on predicting larger granularities of labels.

Conclusions

- MG-SA indeed captures useful phrase information.
- MG-SA promotes the generation of target phrases.
- MG-SA can be applied to many other tasks.