Context-Aware Smoothing for Neural Machine Translation

Kehai Chen[1], Rui Wang[2], Masao Utiyama[2], Eiichiro Sumita[2] and Tiejun Zhao[1]

[1]*Harbin Institute of Technology, China*

[2]*National Institute of Information and Communications Technology, Japan*

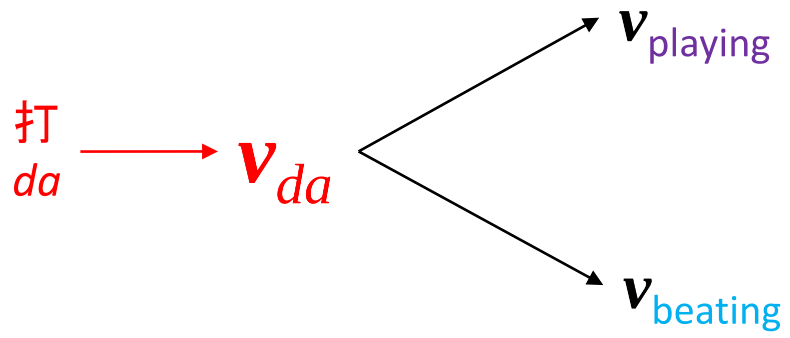
Content

- Motivation
- Related Works
- Context-Aware Representation
- NMT with Context-Aware Smoothing
- Experimental Results
- Conclusion

Motivation-1:polysemy words

Src1	他们	想	通过	打	比赛	来	解决	矛盾
(pinyin)	*tamen*	*xiang*	*tongguo*	*da*	*bisai*	*lai*	*jiejue*	*maodun*
Trg1	*They want to solve the dispute by playing the game*							
Src2	他们	正在	因为	争执	而	打	对方	
(pinyin)	*tamen*	*zhengzai*	*yinwei*	*zhengzhi*	*er*	*da*	*duifang*	
Trg2	*They are beating each other for a dispute*							

Two bilingual parallel sentence pairs

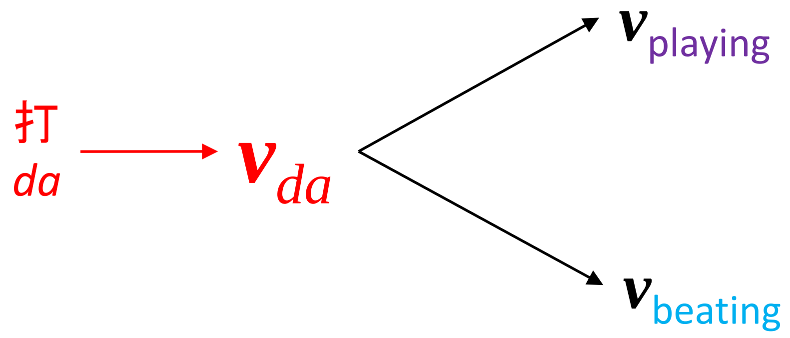


The lexicon semantic depends on its specific context

Motivation-1:polysemy words

Src1	他们	想	通过	打	比赛	来	解决	矛盾
(pinyin)	*tamen*	*xiang*	*tongguo*	*da*	*bisai*	*lai*	*jiejue*	*maodun*
Trg1	*They want to solve the dispute by playing the game*							
Src2	他们	正在	因为	争执	而	打	对方	
(pinyin)	*tamen*	*zhengzai*	*yinwei*	*zhengzhi*	*er*	*da*	*duifang*	
Trg2	*They are beating each other for a dispute*							

Two bilingual parallel sentence pairs



The lexicon semantic depends on its specific context

Motivation-1: Enhancing word representation for polysemy words

Learn specific-sentence word representation $\boldsymbol{v_j}$

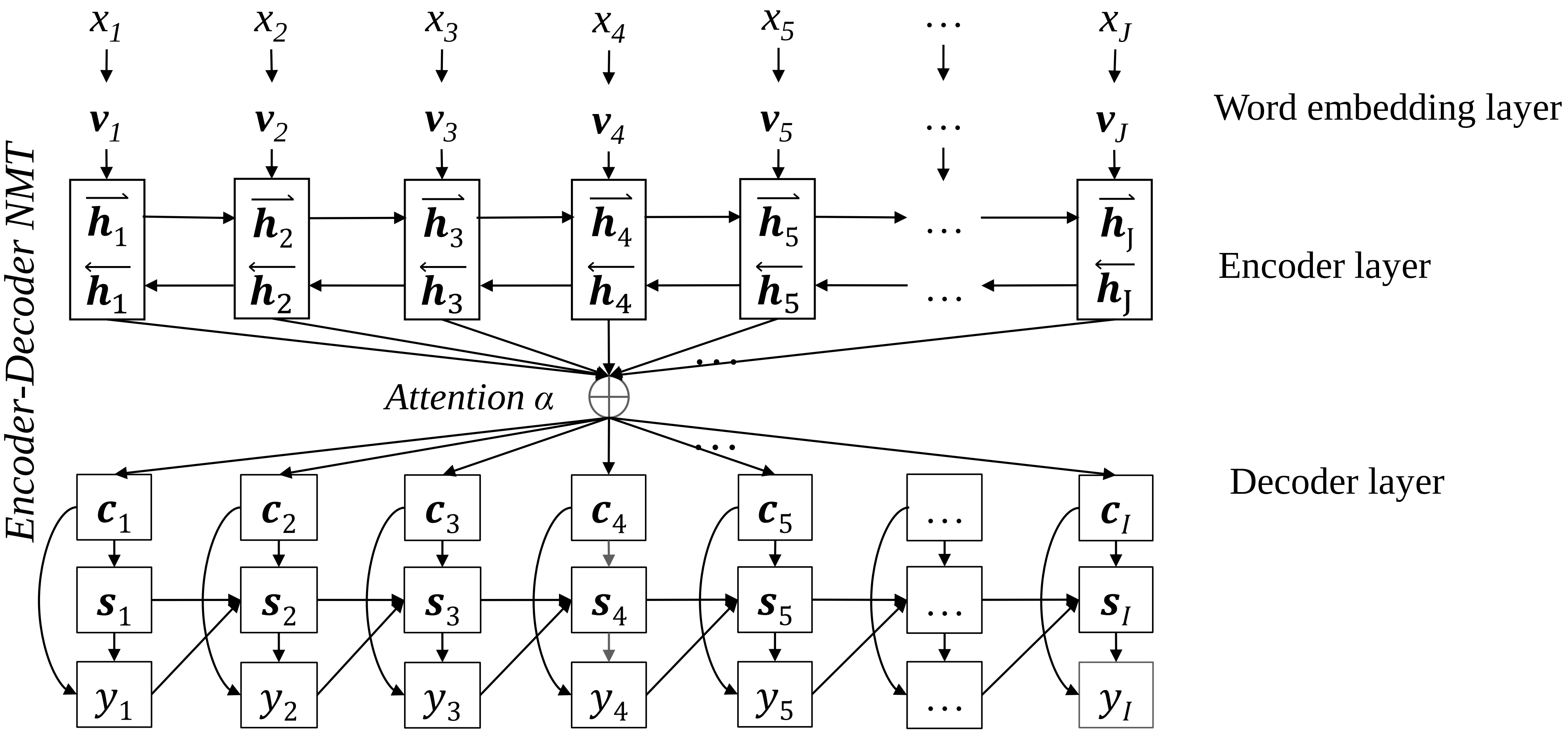
$$\boldsymbol{h_j = f_{enc}(v_j, h_{j-1})}$$

Better source representation

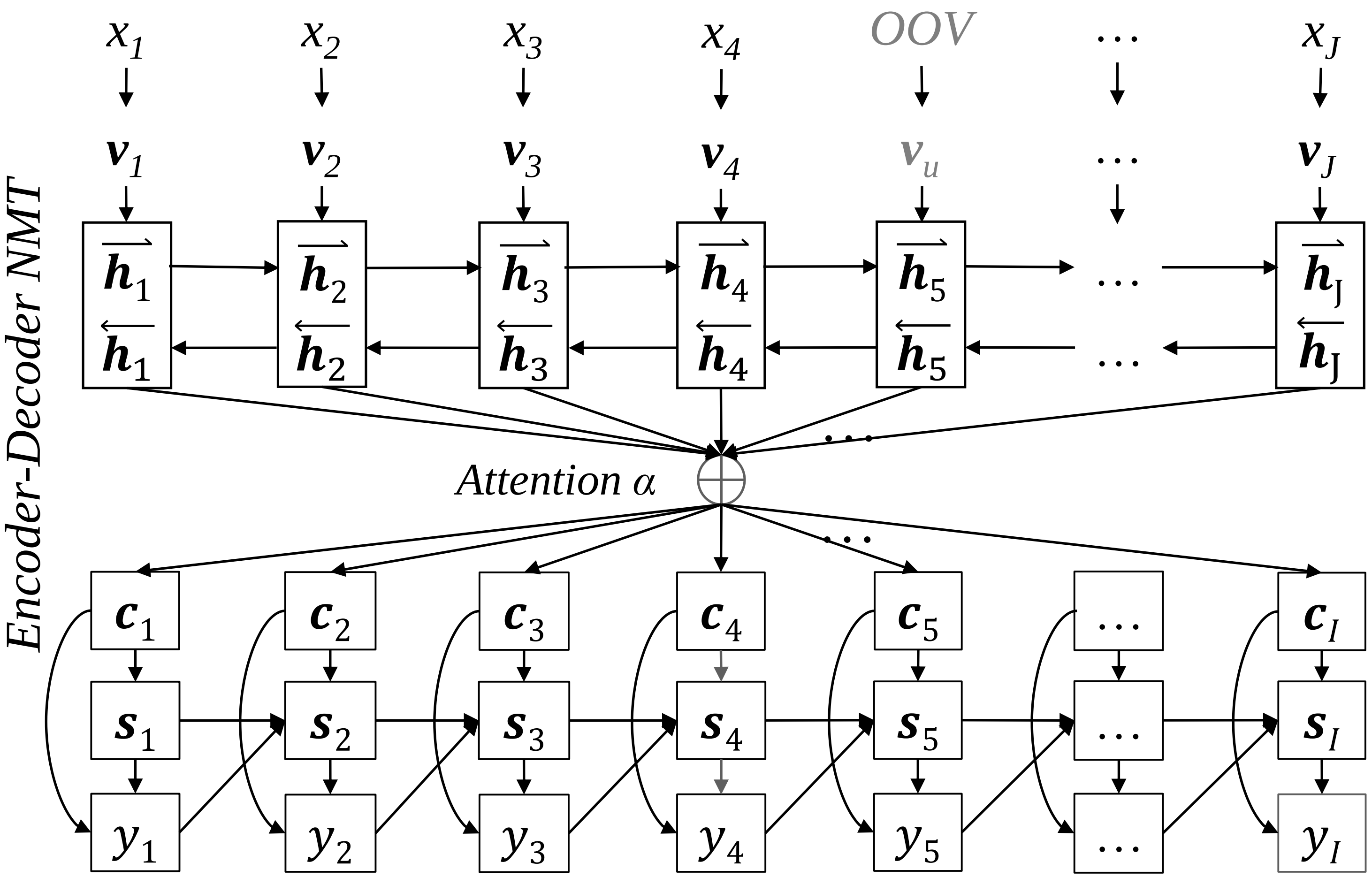
… …

Better target translation

Motivation-2:OOV



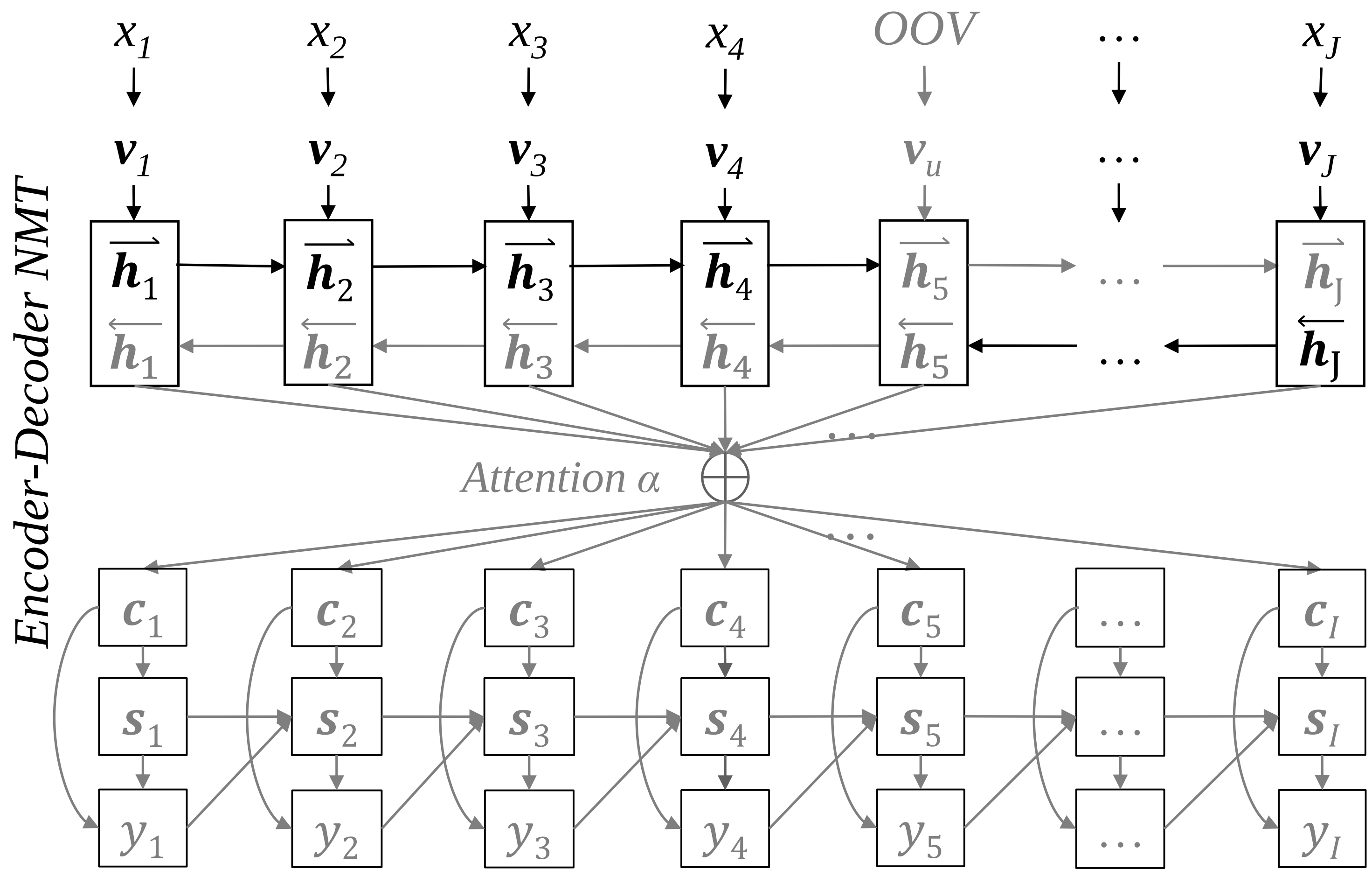
Motivation-2:OOV



- The source sentence includes a OOV

Single vector v_u represents all OOVs

Motivation-2:OOV



These gray parts indicate the parameters of NMT which are affected by the OOV

- The source sentence includes a OOV

Single vector v_u represents all OOVs

- Breaking the structure of sentence;
- Pool source representation;
- … …
- Affecting translation prediction of target word.

Related Works

- Translation Granularity for NMT

  ---Smaller Translation Granularity: Word, Sub-word (BPE), Character for OOV.

  Sennrich et al. (2016), Costa-jussa and Fonollosa (2016), and Li et al. (2016), … …

- Source representation for NMT

  ---RNN or CNN-based Encoder: learning source representation over the sequence of fixed word vectors.

  Bahdanau et al. (2015), Sutskever et al. (2014), … …

Related Works

- Translation Granularity for NMT

  ---Smaller Translation Granularity: Word, Sub-word (BPE), Character for OOV.

  Sennrich et al. (2016), Costa-jussa and Fonollosa (2016), and Li et al. (2016), … …

- Source representation for NMT

  ---RNN or CNN-based Encoder: learning source representation over the sequence of fixed word vectors.

  Bahdanau et al. (2015), Sutskever et al. (2014), … …

- This work focus on enhancing word embedding layer.

  ---Learning a specific-sentence representation for polysemy or OOV word by its context words.

  ---Offering context-aware representation enhances word embedding layer, thereby improving translations (though RNN Encoder can capture word context).

Context-Aware Representation

If there is an OOV "*unk*" (or polysemy word) in the sentence:

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad \textit{unk} \quad x_6 \quad x_7 \quad x_8 \quad x_9$

Context information

$\rightarrow v_u$

$v_1 \quad v_2 \quad v_3 \quad v_4 \quad v_u \quad v_6 \quad v_7 \quad v_8 \quad v_9$

When one understands natural language sentence intuitively, especially including OOV or polysemy word, one often inferences the meaning of these words depending on its context words.

Context-Aware Representation

- We define a context L_j for source word x_j in a fixed size window *2n:*

$$L_j = \underbrace{x_{j-n}, \dots, x_{j-1}}_{\text{Historical } n \text{ words}},\ \underbrace{x_{j+1}, \dots, x_{j+n}}_{\text{Future } n \text{ words}}$$

Context-Aware Representation

- We define a context L_j for source word x_j in a fixed size window *2n:*

$$L_j = \underbrace{x_{j-n}, \ldots, x_{j-1}}_{\text{Historical } n \text{ words}},\ \underbrace{x_{j+1}, \ldots, x_{j+n}}_{\text{Future } n \text{ words}}$$

- Take x_5 as an example, its context L_5 follows (n=2):

$$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7 \quad \cdots \quad x_J$$

$$\boxed{L_5 = x_3, x_4, x_6, x_7}$$

Context-Aware Representation

Feedforward Context-of-Words Model (FCWM)

Output layer:

$$\boldsymbol{V}_{\boldsymbol{L}_{\boldsymbol{j}}} = \sigma(\boldsymbol{W}_{\mathbf{1}}\boldsymbol{L}_{\boldsymbol{j}} + \boldsymbol{b}_{\mathbf{1}})$$

Concatenation:

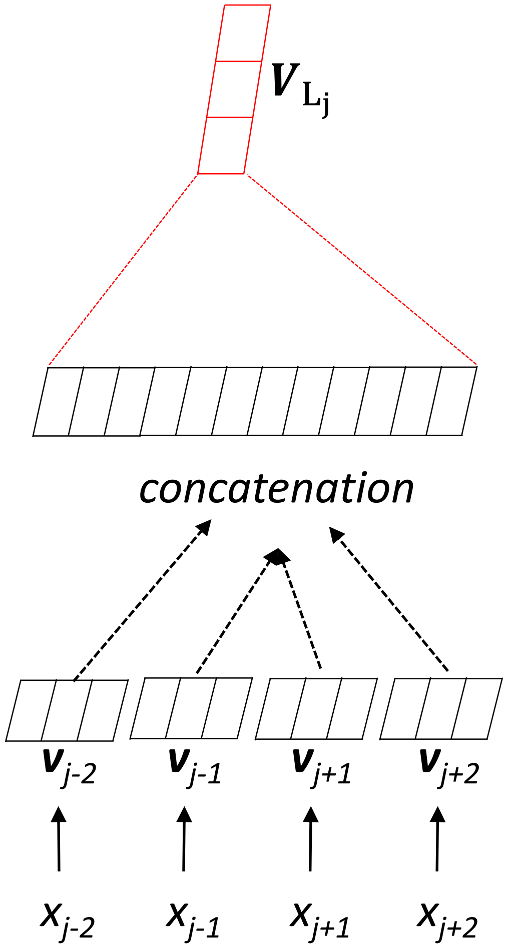
$$\boldsymbol{L}_{\boldsymbol{j}} = [\boldsymbol{v}_{\boldsymbol{j-n}}: \ldots : \boldsymbol{v}_{\boldsymbol{j-1}} : \boldsymbol{v}_{\boldsymbol{j+1}} : \ldots : \boldsymbol{v}_{\boldsymbol{j+n}}]$$

Input layer:

$$\boldsymbol{L}_{\boldsymbol{j}} = \boldsymbol{v}_{\boldsymbol{j-n}}, \ldots, \boldsymbol{v}_{\boldsymbol{j-1}}, \boldsymbol{v}_{\boldsymbol{j+1}}, \ldots, \boldsymbol{v}_{\boldsymbol{j+n}}$$

Context words L_j of x_j:

$$L_j = x_{j-n}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{j+n}$$



$$\boldsymbol{V}_{\boldsymbol{L}_{\boldsymbol{j}}} = \varphi_1\ (L_j;\ \theta_1)$$

Context-Aware Representation

Feedforward Context-of-Words Model (FCWM)

Output layer:

$$V_{L_j} = \sigma(W_1 L_j + b_1)$$

Concatenation:

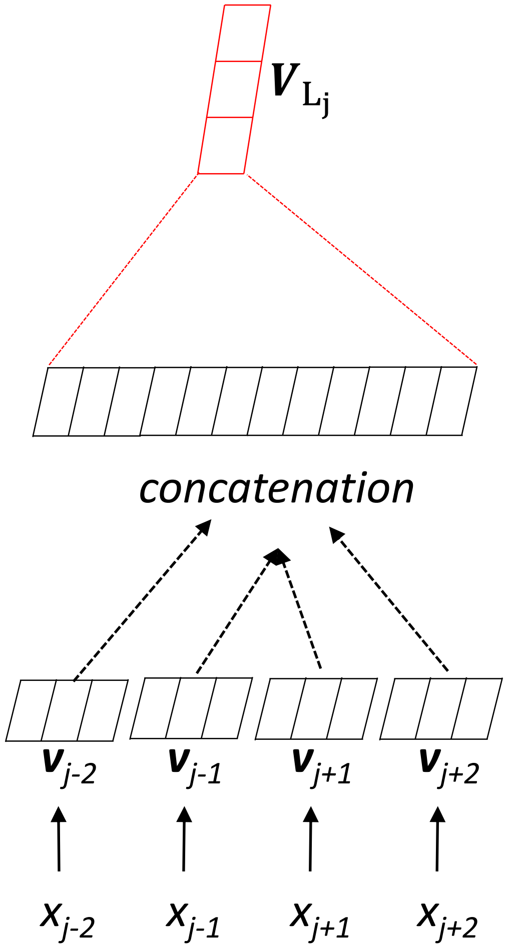
$$L_j = [v_{j-n}: \ldots : v_{j-1}: v_{j+1}: \ldots : v_{j+n}]$$

Input layer:

$$L_j = v_{j-n}, \ldots, v_{j-1}, v_{j+1}, \ldots, v_{j+n}$$

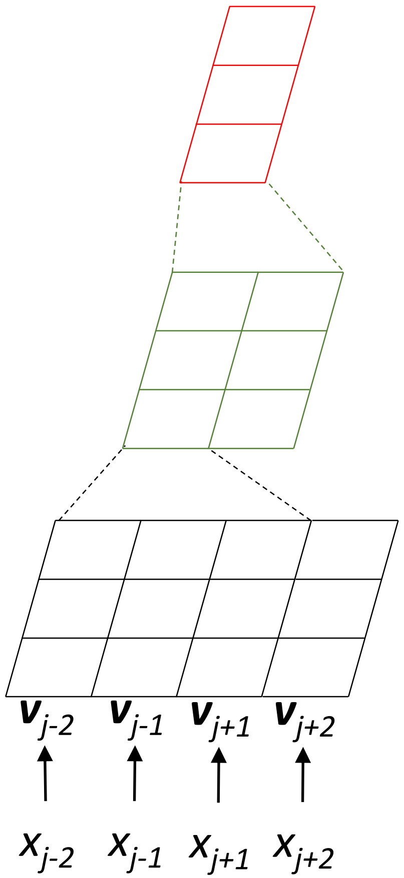
Context words L_j of x_j:

$$L_j = x_{j-n}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{j+n}$$



$$V_{L_j} = \varphi_1 (L_j;\ \theta_1)$$

Convolutional Context-of-Words Model (CCWM)



Non-linear output layer:

$$\mathcal{V}_{\mathcal{L}_j} = \sigma\left(W_3\left(ave\left(\sum_{l=1}^{\frac{2n-k+1}{2}} \mathcal{P}_l\right)\right) + b_3\right)$$

Pooling layer:

$$\mathcal{P} = \max[\mathcal{P}_1, \ldots, \mathcal{P}_{\frac{2n-k+1}{2}}]$$

$$\mathcal{P}_l = \max[\mathcal{L}_{2l-1}, \mathcal{L}_{2l}]$$

Convolution layer:

$$\mathcal{L} = [\mathcal{L}_1, \ldots, \mathcal{L}_{2n-k+1}]$$

$$\mathcal{L}_j = \psi(W_2 \mathcal{M} + b_2)$$

Input layer:

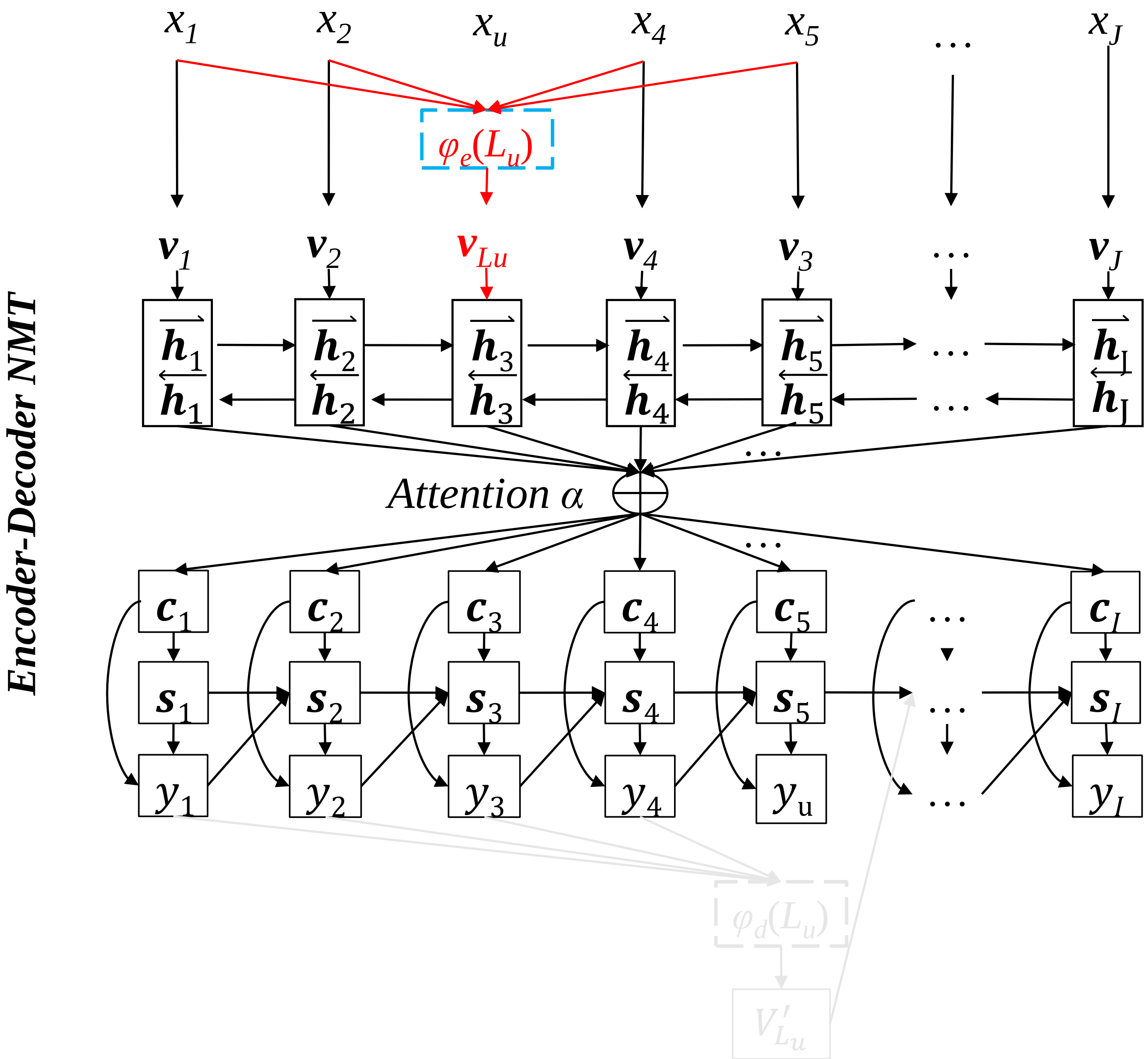
$$\mathcal{M} = [v_{j-n}, \ldots, v_{j-1}, v_{j+1}, \ldots, v_{j+n}]$$

Context words L_j of x_j:

$$L_j = x_{j-n}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{j+n}$$

$$\mathcal{V}_{\mathcal{L}_j} = \varphi_2 (L_j;\ \theta_2)$$

NMT for OOV Smoothing



- ***CARNMT-Enc***

Standard NMT :

$$\boldsymbol{h_j} = f_{enc}(\boldsymbol{v_j}, \boldsymbol{h_{j-1}})$$

This work:

$$\boldsymbol{h_j} = \begin{cases} f_{enc}(\boldsymbol{v_j}, \boldsymbol{h_{j-1}}), & x_j \in V_s \\ f_{enc}(\boldsymbol{\varphi_e}\ (L_{\boldsymbol{x_j}}), \boldsymbol{h_{j-1}}), & x_j \notin V_s \end{cases}$$

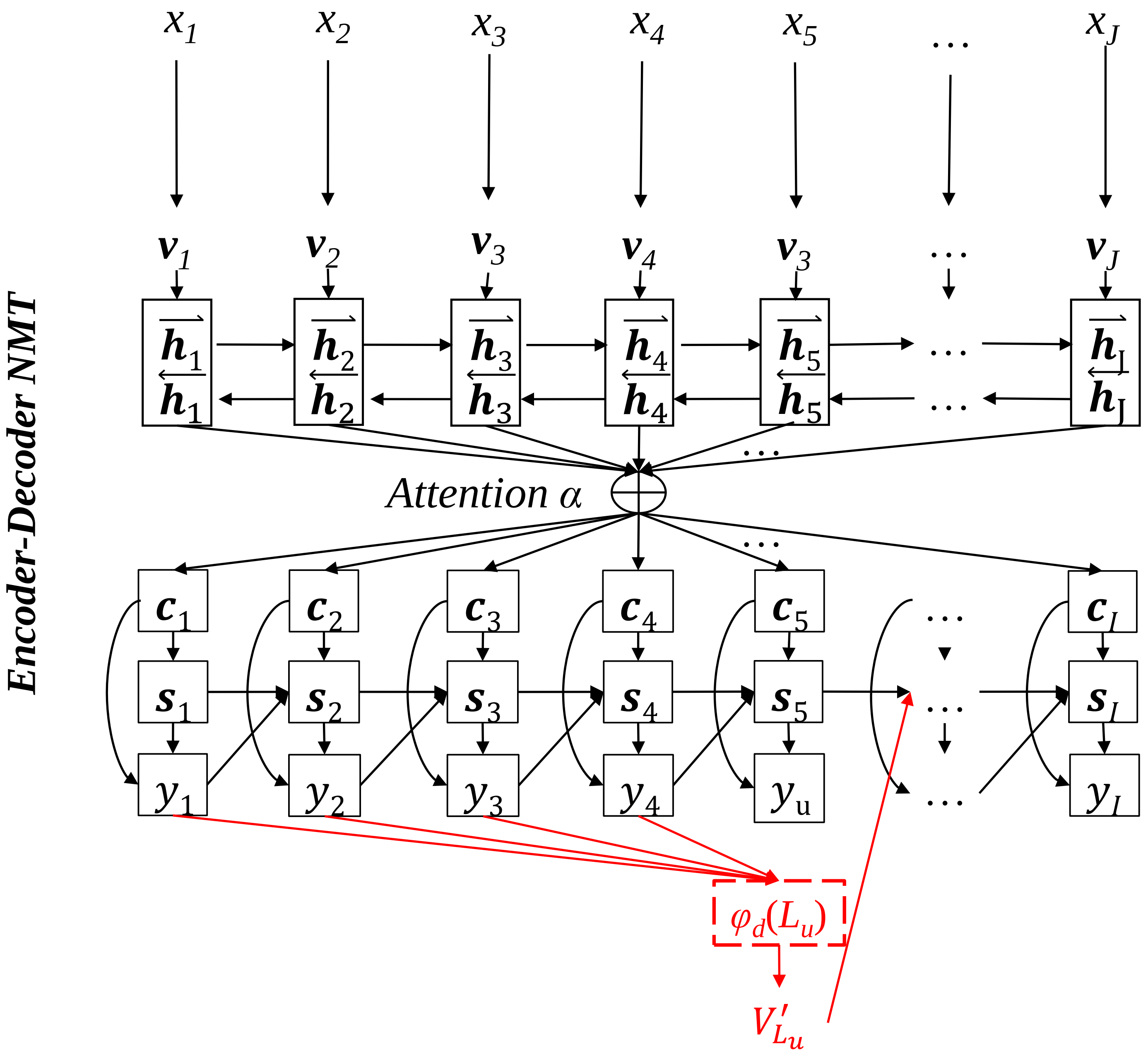
Standard NMT :

$$\boldsymbol{p}(y_i|y_{<i}, x) = g(\boldsymbol{v_{y_{i-1}}}, \boldsymbol{s_i}, \boldsymbol{c_i})$$

This work:

$$\boldsymbol{p}(y_i|y_{<i}, x) = \begin{cases} g(\boldsymbol{v_{y_{i-1}}}, \boldsymbol{s_i}, \boldsymbol{c_i}), & y_{i-1} \in V_t \\ g(\boldsymbol{\varphi_d}\ (L_{y_{i-1}}), \boldsymbol{s_i}, \boldsymbol{c_i}), & y_{i-1} \notin V_t \end{cases}$$

NMT for OOV Smoothing



- ***CARNMT-Dec***

Standard NMT :

$$\boldsymbol{h}_j = f_{enc}(\boldsymbol{v}_j, \boldsymbol{h}_{j-1})$$

This work:

$$\boldsymbol{h}_j = \begin{cases} f_{enc}(\boldsymbol{v}_j, \boldsymbol{h}_{j-1}), & x_j \in V_s \\ f_{enc}(\boldsymbol{\varphi}_e(L_{x_j}), \boldsymbol{h}_{j-1}), & x_j \notin V_s \end{cases}$$

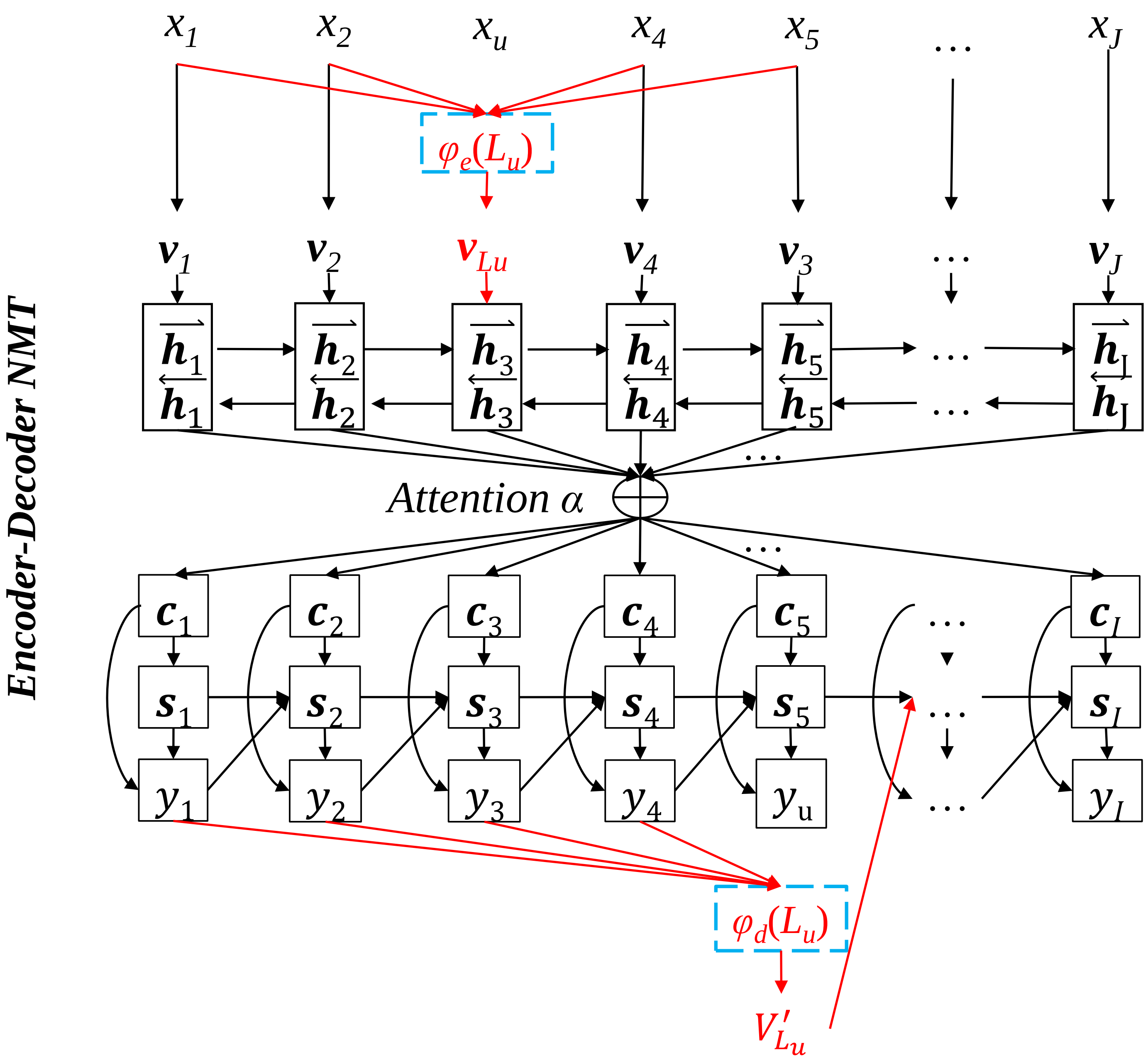
Standard NMT :

$$\boldsymbol{p}(y_i | y_{i<i}, x) = g(\boldsymbol{v}_{\boldsymbol{y}_{i-1}}, \boldsymbol{s}_i, \boldsymbol{c}_i)$$

This work:

$$\boldsymbol{p}(y_i | y_{<i}, x) = \begin{cases} g(\boldsymbol{v}_{\boldsymbol{y}_{i-1}}, \boldsymbol{s}_i, \boldsymbol{c}_i), & y_{i-1} \in V_t \\ g(\boldsymbol{\varphi}_d(L_{\boldsymbol{y}_{i-1}}), \boldsymbol{s}_i, \boldsymbol{c}_i), & y_{i-1} \notin V_t \end{cases}$$

NMT for OOV Smoothing



- ***CARNMT-Both***

Standard NMT :

$$\boldsymbol{h_j} = f_{enc}(\boldsymbol{v_j}, \boldsymbol{h_{j-1}})$$

This work:

$$\boldsymbol{h_j} = \begin{cases} f_{enc}(\boldsymbol{v_j}, \boldsymbol{h_{j-1}}), & x_j \in V_s \\ f_{enc}(\boldsymbol{\varphi}_e\,(L_{\boldsymbol{x_j}}), \boldsymbol{h_{j-1}}), & x_j \notin V_s \end{cases}$$

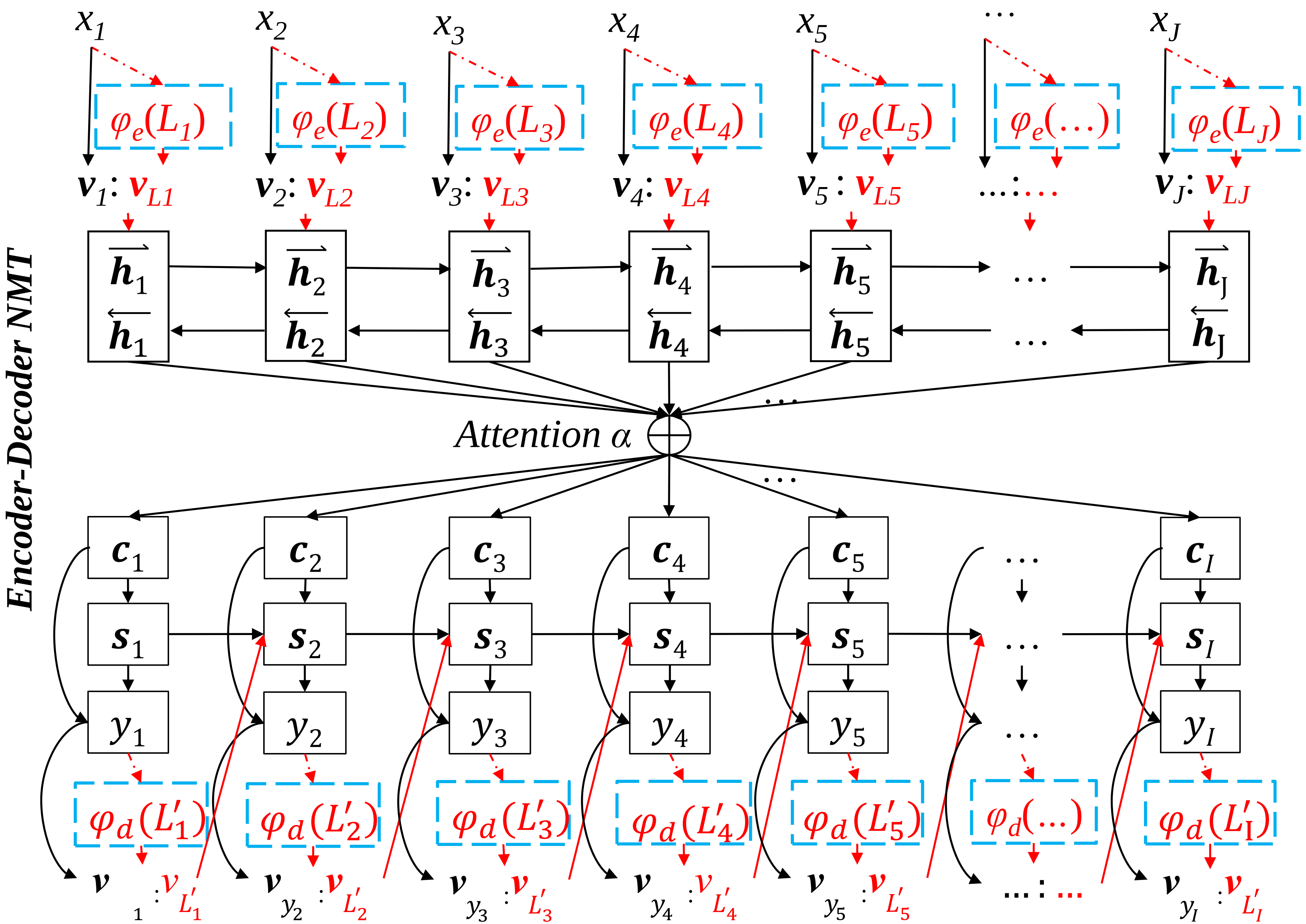
Standard NMT :

$$\boldsymbol{p}(y_i|y_{i<i}, x) = g(\boldsymbol{v}_{\boldsymbol{y_{i-1}}}, \boldsymbol{s_i}, \boldsymbol{c_i})$$

This work:

$$\boldsymbol{p}(y_i|y_{<i}, x) = \begin{cases} g(\boldsymbol{v}_{\boldsymbol{y_{i-1}}}, \boldsymbol{s_i}, \boldsymbol{c_i}), & y_{i-1} \in V_t \\ g(\boldsymbol{\varphi}_d\,(L_{\boldsymbol{y_{i-1}}}), \boldsymbol{s_i}, \boldsymbol{c_i}), & y_{i-1} \notin V_t \end{cases}$$

NMT for Smoothing all words



- ***CARNMT-ALL***

Standard NMT :

$$\boldsymbol{h_j} = f_{enc}(\boldsymbol{v_j}, \boldsymbol{h_{j-1}})$$

This work:

$$\boldsymbol{h_j} = f_{enc}(\boldsymbol{\varphi_e}\,(L_{\boldsymbol{x_j}}), \boldsymbol{h_{j-1}})$$

Standard NMT :

$$\boldsymbol{p}(y_i|y_{i<i}, x) = g(\boldsymbol{v_{y_{i-1}}}, \boldsymbol{s_i}, \boldsymbol{c_i})$$

This work:

$$\boldsymbol{p}(y_i|y_{<i}, x) = g(\boldsymbol{\varphi_d}\,(L_{\boldsymbol{y_{i-1}}}), \boldsymbol{s_i}, \boldsymbol{c_i})$$

Experimental Settings

- Training data includes 1.42M Chinese-to-English parallel sentence pairs from *LDC corpus*.
- The NIST 2002 (MT02) and NIST 2003-2008 (MT03-08) datasets are as validation set and test sets, respectively. The Case-insensitive *4*-gram NIST BLEU score (Papineni et al., 2002) is as evaluation metric.
- Vocab is 30k; Sentence length is 80; Mini-batch size 80; Word embedding dim is 620; Hidden layer dim is 1000; Dropout on the all layers; Optimizer is Adadelta.
- The baseline includes: Standard Attentional NMT (Bahdanau et al., 2014); Subword-based NMT (Sennrich et al., 2016); Character-based NMT (Costa-jussa and Fonollosa, 2016); Replacing unk with similarity semantic in vocabulary words (Li et al., 2016).

Experimental Results

- Results for Chinese-to-English Translation Task

System	Dev (MT02)	MT03	MT04	MT05	MT06	MT08	AVG
Moses	33.15	31.02	33.78	30.33	29.62	23.53	29.66
Bahdanau et al. (2015)	36.42	34.22	37.11	33.02	32.69	25.38	32.48
Sennrich et al. (2016)	36.89	35.39	38.24	33.73	32.74	26.22	33.26
Costa-jussà and Fonollosa (2016)	35.98	34.93	37.56	33.24	32.32	26.02	32.81
Li et al. (2016)	36.96	35.78	38.42	34.02	33.14	26.36	33.54
CARNMT-Encoder (FCWM)	36.78	35.56**	38.14*	33.69	33.13	26.16*	33.34
CARNMT-Decoder (FCWM)	36.67	34.65	37.60	33.26	33.01	26.15*	32.93
CARNMT-Both (FCWM)	37.36	35.43**	38.34**	33.43	**33.47**	**26.86****	33.50
ALLSmooth (FCWM)	37.71	35.73**	**38.53****	33.91*	**33.53***	**27.18****	**33.78**
CARNMT-Encoder (CCWM)	37.12	35.64**	38.14*	33.49	33.26*	26.57**	33.42
CARNMT-Decoder (CCWM)	36.33	34.56	37.43	33.24	32.96	25.86	32.81
CARNMT-Both (CCWM)	37.56	**35.83****	**38.52****	33.73	**33.37****	**27.06****	**33.70**
ALLSmooth (CCWM)	37.69	**36.23****	**38.89****	**34.69****	**33.83****	**27.94**†	**34.32**

- Moses *VS* NMT ------> Strong baselines
- CARNMT-Enc/Dec *VS* Bahdanau et al. (2015) ------> Our method can effectively smooth the negative effect (Motivation 1)
- CARNMT-Both *VS* CARNMT-Enc/Dec ------> Source-side smoothing is orthogonal with target-side smoothing (Motivation 1)
- ALLSmooth *VS* CARNMT-Both ------> In-vocabulary smoothing is beneficial for NMT (Motivation 2)

Experimental Results

- Results for Chinese-to-English Translation Task

System	**Dev (MT02)**	**MT03**	**MT04**	**MT05**	**MT06**	**MT08**	**AVG**
Moses	33.15	31.02	33.78	30.33	29.62	23.53	29.66
Bahdanau et al. (2015)	36.42	34.22	37.11	33.02	32.69	25.38	32.48
Sennrich et al. (2016)	36.89	35.39	38.24	33.73	32.74	26.22	33.26
Costa-jussà and Fonollosa (2016)	35.98	34.93	37.56	33.24	32.32	26.02	32.81
Li et al. (2016)	36.96	35.78	38.42	34.02	33.14	26.36	33.54
CARNMT-Encoder (FCWM)	36.78	35.56**	38.14*	33.69	33.13	26.16*	33.34
CARNMT-Decoder (FCWM)	36.67	34.65	37.60	33.26	33.01	26.15*	32.93
CARNMT-Both (FCWM)	37.36	35.43**	38.34**	33.43	**33.47**	**26.86****	33.50
ALLSmooth (FCWM)	37.71	35.73**	**38.53****	33.91*	**33.53***	**27.18****	**33.78**
CARNMT-Encoder (CCWM)	37.12	35.64**	38.14*	33.49	33.26*	26.57**	33.42
CARNMT-Decoder (CCWM)	36.33	34.56	37.43	33.24	32.96	25.86	32.81
CARNMT-Both (CCWM)	37.56	**35.83****	**38.52****	33.73	**33.37****	**27.06****	**33.70**
ALLSmooth (CCWM)	37.69	**36.23****	**38.89****	**34.69****	**33.83****	**27.94**†	**34.32**

- CARNMT-Enc/Dec *VS* Bahdanau et al. (2015)
  Our smooth method can relieve the negative effect of OOV effectively, as in Motivation 2
- CARNMT-Both *VS* CARNMT-Enc/Dec ------> Source-side smoothing is orthogonal with target-side smoothing (Motivation 1)
- ALLSmooth *VS* CARNMT-Both ------> In-vocabulary smoothing is beneficial for NMT (Motivation 2)

Experimental Results

- Results for Chinese-to-English Translation Task

System	Dev (MT02)	MT03	MT04	MT05	MT06	MT08	AVG
Moses	33.15	31.02	33.78	30.33	29.62	23.53	29.66
Bahdanau et al. (2015)	36.42	34.22	37.11	33.02	32.69	25.38	32.48
Sennrich et al. (2016)	36.89	35.39	38.24	33.73	32.74	26.22	33.26
Costa-jussà and Fonollosa (2016)	35.98	34.93	37.56	33.24	32.32	26.02	32.81
Li et al. (2016)	36.96	35.78	38.42	34.02	33.14	26.36	33.54
CARNMT-Encoder (FCWM)	36.78	35.56**	38.14*	33.69	33.13	26.16*	33.34
CARNMT-Decoder (FCWM)	36.67	34.65	37.60	33.26	33.01	26.15*	32.93
CARNMT-Both (FCWM)	37.36	35.43**	38.34**	33.43	**33.47**	**26.86****	33.50
ALLSmooth (FCWM)	37.71	35.73**	**38.53****	33.91*	**33.53***	**27.18****	**33.78**
CARNMT-Encoder (CCWM)	37.12	35.64**	38.14*	33.49	33.26*	26.57**	33.42
CARNMT-Decoder (CCWM)	36.33	34.56	37.43	33.24	32.96	25.86	32.81
CARNMT-Both (CCWM)	37.56	**35.83****	**38.52****	33.73	**33.37****	**27.06****	**33.70**
ALLSmooth (CCWM)	37.69	**36.23****	**38.89****	**34.69****	**33.83****	**27.94**†	**34.32**

- CARNMT-Both *VS* CARNMT-Enc/Dec

  Source-side smoothing is orthogonal with target-side smoothing

Experimental Results

- Results for Chinese-to-English Translation Task

System	Dev (MT02)	MT03	MT04	MT05	MT06	MT08	AVG
Moses	33.15	31.02	33.78	30.33	29.62	23.53	29.66
Bahdanau et al. (2015)	36.42	34.22	37.11	33.02	32.69	25.38	32.48
Sennrich et al. (2016)	36.89	35.39	38.24	33.73	32.74	26.22	33.26
Costa-jussà and Fonollosa (2016)	35.98	34.93	37.56	33.24	32.32	26.02	32.81
Li et al. (2016)	36.96	35.78	38.42	34.02	33.14	26.36	33.54
CARNMT-Encoder (FCWM)	36.78	35.56**	38.14*	33.69	33.13	26.16*	33.34
CARNMT-Decoder (FCWM)	36.67	34.65	37.60	33.26	33.01	26.15*	32.93
CARNMT-Both (FCWM)	37.36	35.43**	38.34**	33.43	**33.47**	**26.86****	33.50
ALLSmooth (FCWM)	37.71	35.73**	**38.53****	33.91*	**33.53***	**27.18****	**33.78**
CARNMT-Encoder (CCWM)	37.12	35.64**	38.14*	33.49	33.26*	26.57**	33.42
CARNMT-Decoder (CCWM)	36.33	34.56	37.43	33.24	32.96	25.86	32.81
CARNMT-Both (CCWM)	37.56	**35.83****	**38.52****	33.73	**33.37****	**27.06****	**33.70**
ALLSmooth (CCWM)	37.69	**36.23****	**38.89****	**34.69****	**33.83****	**27.94**†**	**34.32**

- ALLSmooth *VS* CARNMT-Both

  In-vocabulary smoothing is also beneficial for NMT (Motivation 1)

Experimental Results

- Results for Chinese-to-English Translation Task

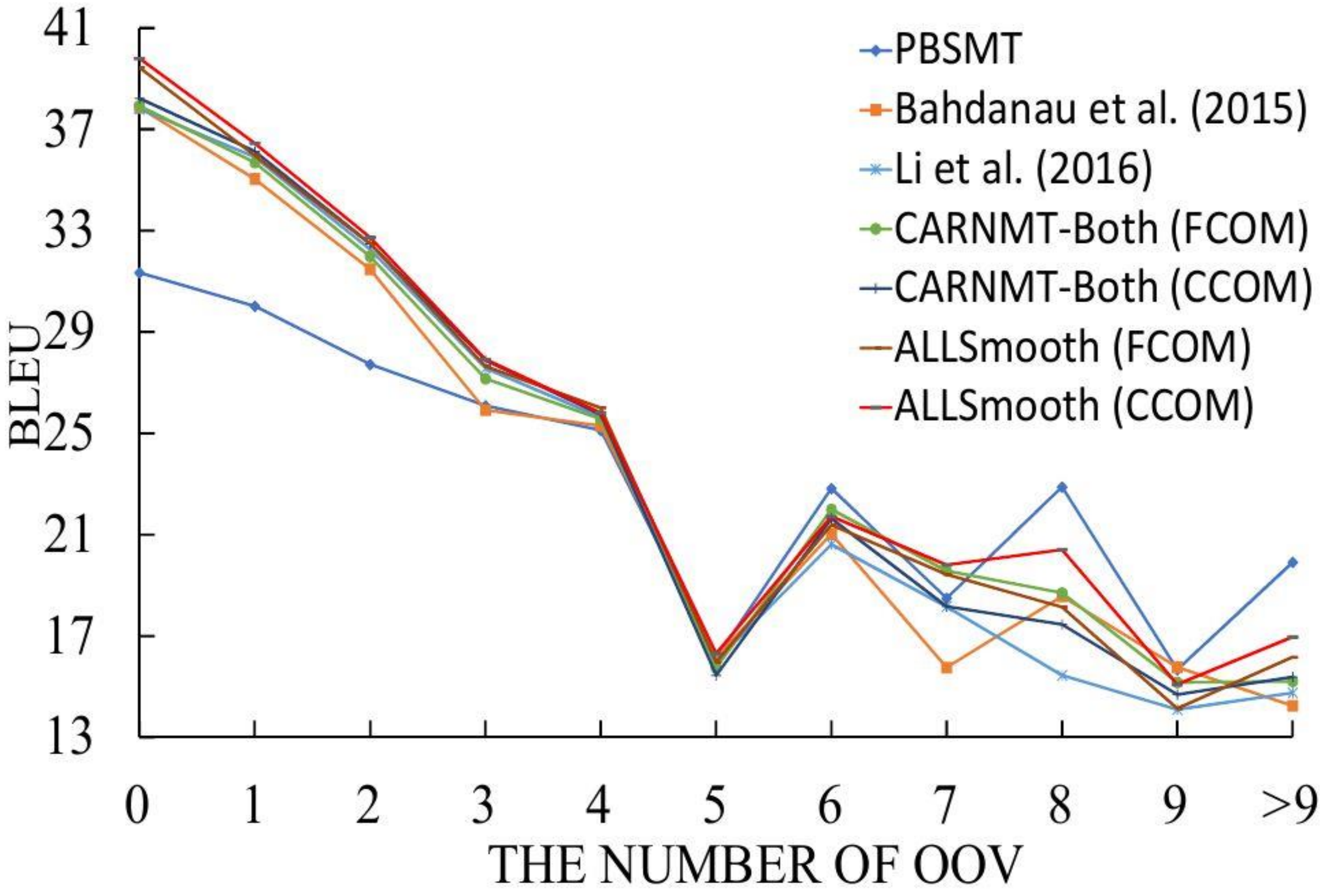
System	Dev (MT02)	MT03	MT04	MT05	MT06	MT08	AVG
Moses	33.15	31.02	33.78	30.33	29.62	23.53	29.66
Bahdanau et al. (2015)	36.42	34.22	37.11	33.02	32.69	25.38	32.48
Sennrich et al. (2016)	36.89	35.39	38.24	33.73	32.74	26.22	33.26
Costa-jussà and Fonollosa (2016)	35.98	34.93	37.56	33.24	32.32	26.02	32.81
Li et al. (2016)	36.96	35.78	38.42	34.02	33.14	26.36	33.54
CARNMT-Encoder (FCWM)	36.78	35.56**	38.14*	33.69	33.13	26.16*	33.34
CARNMT-Decoder (FCWM)	36.67	34.65	37.60	33.26	33.01	26.15*	32.93
CARNMT-Both (FCWM)	37.36	35.43**	38.34**	33.43	**33.47**	**26.86****	33.50
ALLSmooth (FCWM)	37.71	35.73**	**38.53****	33.91*	**33.53***	**27.18****	**33.78**
CARNMT-Encoder (CCWM)	37.12	35.64**	38.14*	33.49	33.26*	26.57**	33.42
CARNMT-Decoder (CCWM)	36.33	34.56	37.43	33.24	32.96	25.86	32.81
CARNMT-Both (CCWM)	37.56	**35.83****	**38.52****	33.73	**33.37****	**27.06****	**33.70**
ALLSmooth (CCWM)	37.69	**36.23****	**38.89****	**34.69****	**33.83****	**27.94**†	**34.32**

- FCWM *VS* CCWM

The CCWM learns the context semantic representation directly for smoothing word vector, while the FCWM predicts semantic representation of word depending on its context.

Experimental Results

- Translation Qualities for Sentences with Different Numbers of OOV

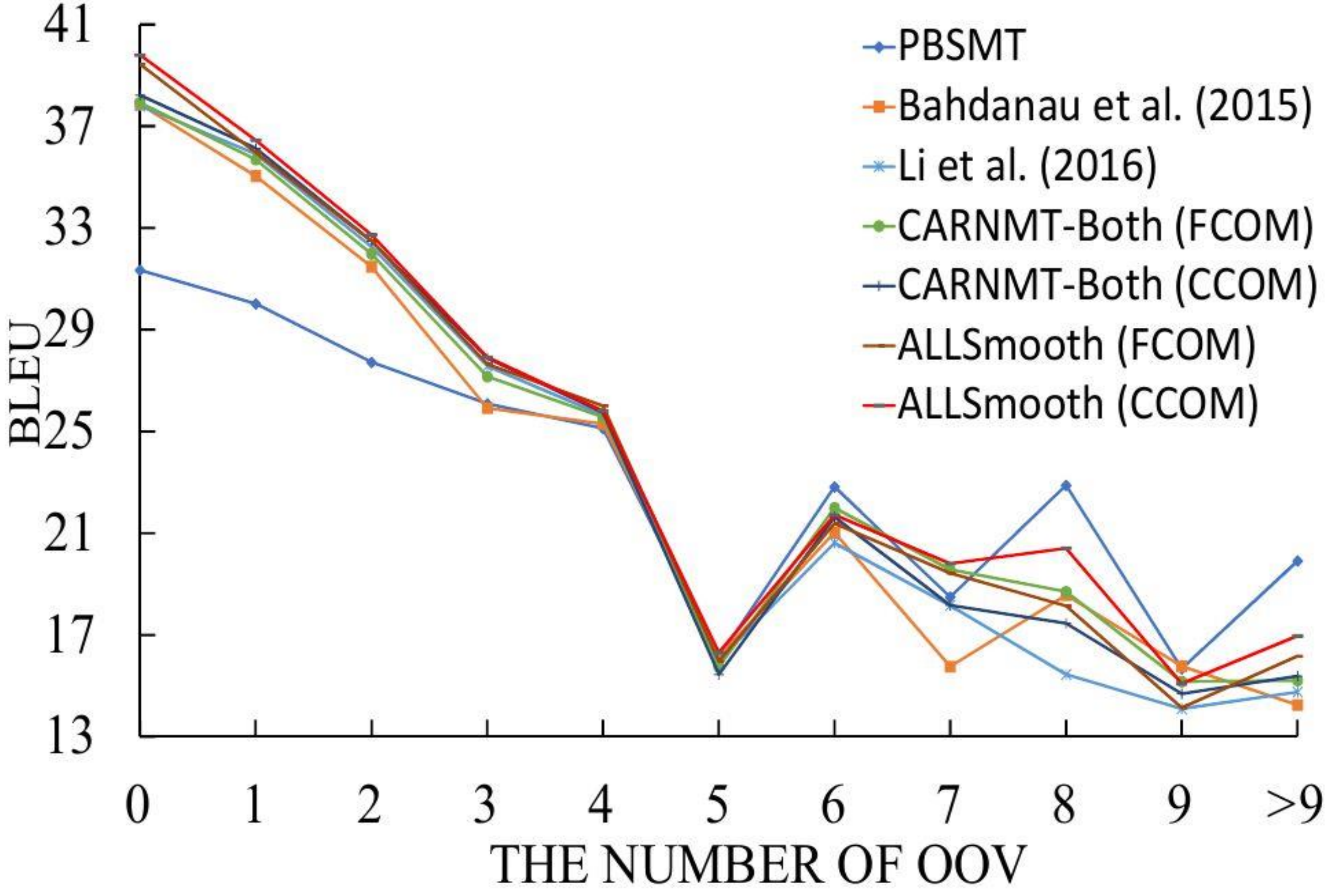


2306	1827	1121	678	391	215	123	59	37	24	29	The number of sentences

- The number of OOV = 0
  - ALLSmooth is better than the baseline Bahdanau et al. (2015).
  - Both of CARNMT-Enc/Dec are similar to baseline Bahdanau et al. (2015).
- With the increasing in the number of OOVs
  - The gap between our methods and other methods (except PBSMT) become larger, especially when more than five.
- When the number of OOV is more than seven
  - PBSMT is better than all NMT models

Experimental Results

- Translation Qualities for Sentences with Different Numbers of OOV

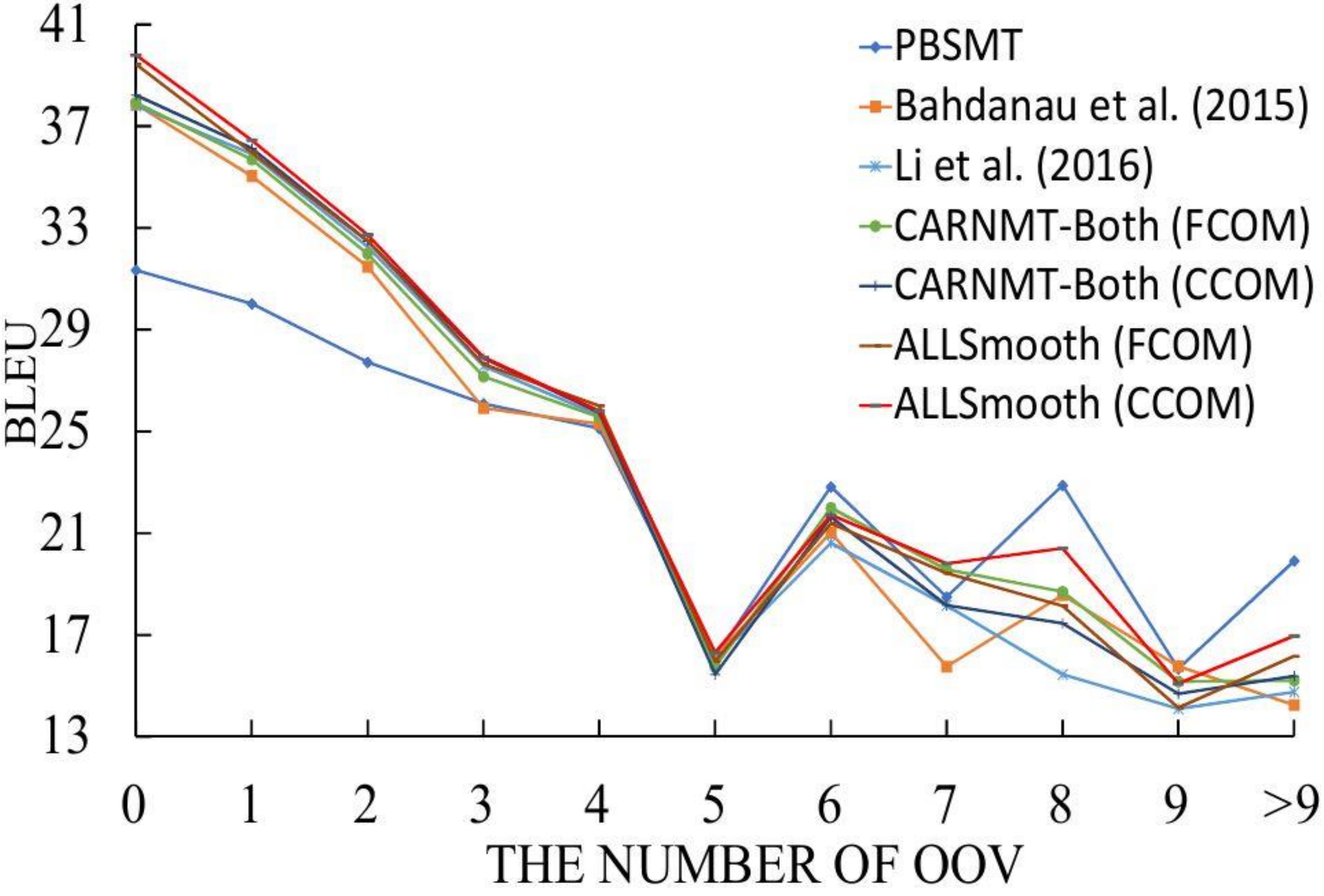


- The number of OOV = 0
  - ALLSmooth is better than the baseline Bahdanau et al. (2015).
  - Both of CARNMT-Enc/Dec are similar to baseline Bahdanau et al. (2015).
- With the increasing in the number of OOVs
  - The gap between our methods and other methods (except PBSMT) become larger, especially when more than five.
- When the number of OOV is more than seven
  - PBSMT is better than all NMT models

2306	1827	1121	678	391	215	123	59	37	24	29	The number of sentences

Experimental Results

- Translation Qualities for Sentences with Different Numbers of OOV

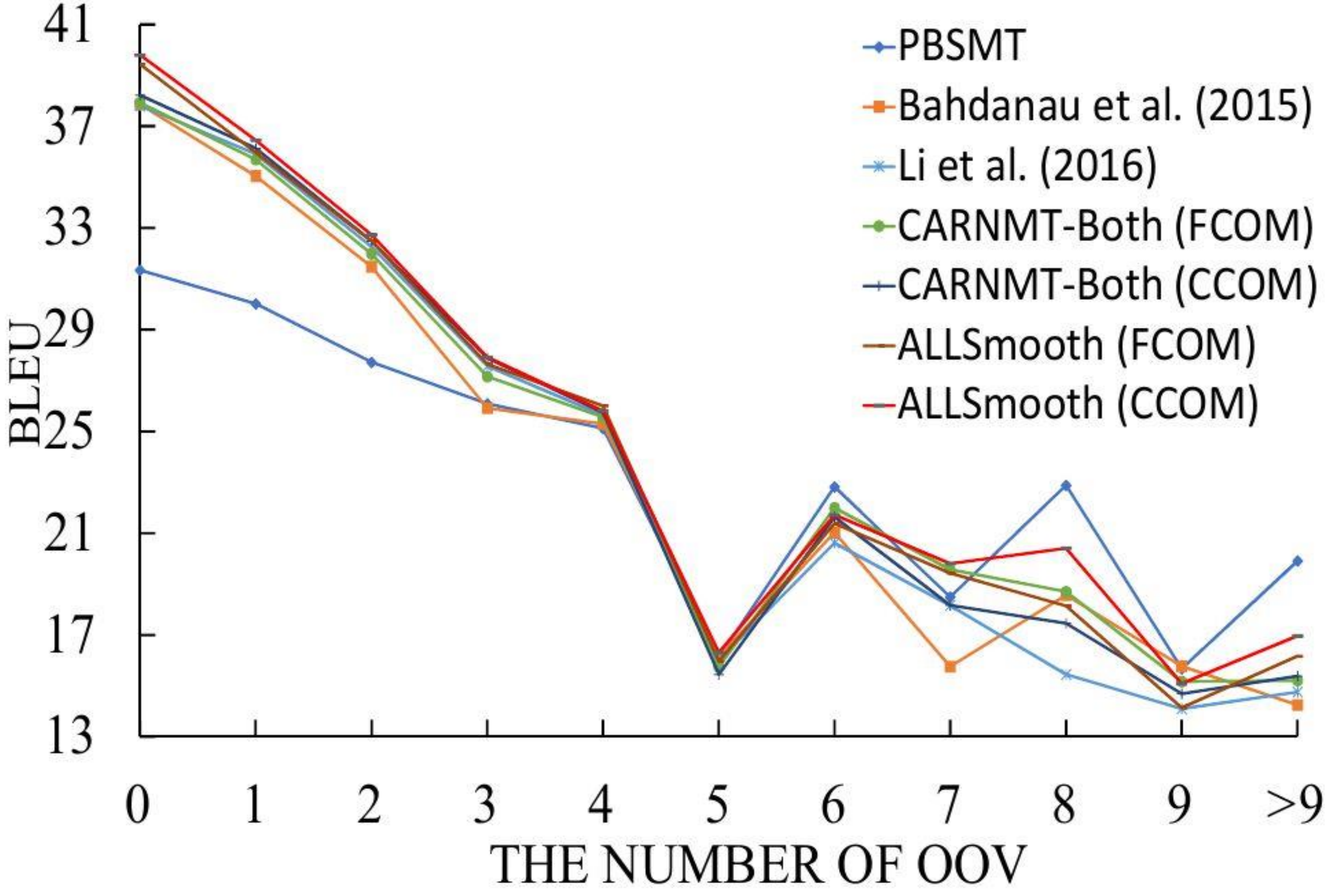


- The number of OOV = 0
  - ALLSmooth is better than the baseline Bahdanau et al. (2015).
  - Both of CARNMT-Enc/Dec are similar to baseline Bahdanau et al. (2015).
- With the increasing in the number of OOVs
  - The gap between our methods and other methods (except PBSMT) become larger, especially when more than five.
- When the number of OOV is more than seven
  - PBSMT is better than all NMT models

2306	1827	1121	678	391	215	123	59	37	24	29	The number of sentences

Experimental Results

- Translation Qualities for Sentences with Different Numbers of OOV



- The number of OOV = 0
  - ALLSmooth is better than the baseline Bahdanau et al. (2015).
  - Both of CARNMT-Enc/Dec are similar to baseline Bahdanau et al. (2015).
- With the increasing in the number of OOVs
  - The gap between our methods and other methods (except PBSMT) become larger, especially when more than five.
- When the number of OOV is more than seven
  - PBSMT is better than all NMT models

2306	**1827**	**1121**	**678**	**391**	**215**	**123**	**59**	**37**	**24**	**29**	The number of sentences

Experimental Results

- Translation Qualities for Sentences with Different Numbers of OOV

SRC: 用好 这个 战略 机遇期 (OOV) , 力争 有所 作为 , 必须 把 发展 科学 技术 放在 更加 重要 , 更加 突出 的 位置
(**pinyin**) *yonghao zhege zhanlue jiyuqi , lizheng yousuo zuowei , bixu ba fazhan kexue jishu fangzai gengjia zhongyao , gengjia tuchu de wiezhi*

Bahdanauet al.(2015) : to make good use of this strategy , we should strive for the development of science and technology , and must put the development of science and technology into an even more important and prominent position

This work: in making good use of this strategic plan and striving to accomplish something , it is necessary to place the development of science and technology in a more important and more prominent position

Ref: to well use this strategic period of opportunity and strive to accomplish some achievments , the development of science and technology should be placed in a more prior and prominent position

- The negative effect of OOV exists in NMT
  - The OOV "*jiyuqi*" itself is not translated.
  - The phrase "*lizheng yousuo zuowei*" (the red part in English) is not translated.
- Smoothing the negative effect of OOV
  - Obtaining the translation "*striving to accomplish something*" of "*lizheng yousuo zuowei*".

Conclusion

- Experimental results showed that the negative effect of OOV decreased the translation performance of NMT, and the existing RNN encoder can not adequately address the problem.
- The learned CAR was integrated into the Encoder to smooth word representation, and thus enhanced the Decoder of NMT.
- Experimental results showed that the proposed method can greatly alleviate the negative effect of OOV and enhance word representation of in-vocabulary words, thus improving the translations.

Q&A
Thanks