The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Syntax-Directed Attention for Neural Machine Translation

Kehai Chen,[1*] **Rui Wang,**[2†] **Masao Utiyama,**[2] **Eiichiro Sumita,**[2] **Tiejun Zhao**[1]
[1]Harbin Institute of Technology, Harbin, China
[2]National Institute of Information and Communications Technology, Kyoto, Japan
{khchen, tjzhao}@hit.edu.cn, {wangrui, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

Attention mechanism, including global attention and local attention, plays a key role in neural machine translation (**NMT**). Global attention attends to all source words for word prediction. In comparison, local attention selectively looks at fixed-window source words. However, alignment weights for the current target word often decrease to the left and right by linear distance centering on the aligned source position and neglect syntax distance constraints. In this paper, we extend the local attention with syntax-distance constraint, which focuses on syntactically related source words with the predicted target word to learning a more effective context vector for predicting translation. Moreover, we further propose a double context NMT architecture, which consists of a global context vector and a syntax-directed context vector from the global attention, to provide more translation performance for NMT from source representation. The experiments on the large-scale Chinese-to-English and English-to-German translation tasks show that the proposed approach achieves a substantial and significant improvement over the baseline system.

1 Introduction

Recent works of neural machine translation (**NMT**) have been proposed to adopt the encoder-decoder framework (Kalchbrenner and Blunsom 2013; Cho et al. 2014; Sutskever, Vinyals, and Le 2014), which employs a recurrent neural network (**RNN**) encoder to represent source sentence and a RNN decoder to generate target translation word by word. Especially, the NMT with an *attention mechanism* (called as global attention) is proposed to acquire source sentence context dynamically at each decoding step, thus improving the performance of NMT (Bahdanau, Cho, and Bengio 2015). The global attention is further refined into a local attention (Luong, Pham, and Manning 2015), which selectively looks at fixed-window source context at each decoding step, thus demonstrating its effectiveness on WMT translation tasks between English and German in both directions.

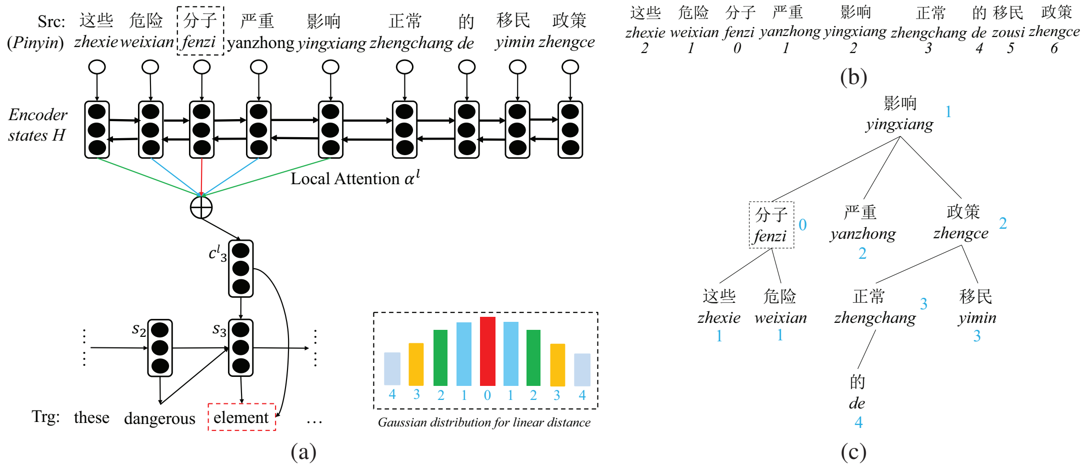
Specifically, the local attention first predicts a single aligned source position p_i for the current time-step i. The decoder focuses on the fixed-window encoder states centered around the source position p_i, and compute a context vector $\boldsymbol{c}_i^l$ by alignment weights α^l for predicting current target word. Figure 1(a) shows a Chinese-to-English NMT model with the local attention, and its contextual window is set to five. When the aligned source word is "*fenzi*", the local attention focuses on source words {"*zhexie*", "*weixian*", "*fenzi*", "*yanzhong*", "*yingxiang*"} in the window to compute its context vector. Meanwhile, the local attention is to obtain the positions of five encoder states by Gaussian distribution, which penalty their alignment weights according to the distance with word "*fenzi*". For example, the syntax distances of these five source words are {*2*, *1*, *0*, *1*, *2*} in contextual window, as shown in Figure 1(b). In other words, the greater the distance from the aligned word in the window is, the smaller the source words in the window to the context vector would contribute. In spite of its success, the local attention is to encode source context and compute a local context vector by linear distance centered around current aligned source position. It does not take syntax distance constraints into account.

Figure 1(c) shows the dependency tree of the Chinese sentence in Figure 1(b). Support the word "*fenzi*$_0$" as the aligned source word, its syntax-distance neighbor window is {"*zhexie*$_1$", "*weixian*$_1$", "*fenzi*$_0$", "*yingxiang*$_1$", "*yanzhong*$_2$", "*zhengce*$_2$"} , where the footnote of a word is its syntax-distance with the central word. In comparison, its local neighbor window is {"*zhexie*", "*weixian*", "*yanzhong*", "*yingxiang*", "*zhengchang*"} based on linear distance. Note that the "*zhengce*" is very informative for the correct translation, but it is far away from "*fenzi*" such that it is not easy to be focused by the local attention. Besides, the syntax distances of "*yanzhong*" and "*yingxiang*" are two and one, but the linear distances are one and two. This means that the "*yingxiang*" is syntactically more relevant to the "*fenzi*" than "*yingxiang*". However, the existing *attention mechanism*, including the global or local attention, does not allow NMT to distinguish syntax distance constraint from source representation.

In this paper, we extend the local attention with a novel syntax-distance constraint, to capture syntax related source words with the predicted target word. Following the dependency tree of a source sentence, each source word has a syntax-distance constraint mask, which denotes its syntax

[*] Kehai Chen was an internship research fellow at NICT when conducting this work.

[†] Corresponding author.



Src: 这些 危险 分子 严重 影响 正常 的 移民 政策
(Pinyin) zhexie weixian fenzi yanzhong yingxiang zhengchang de yimin zhengce
Encoder states H
Local Attention α^l
c^l_3
s_2
s_3
Trg: these dangerous element …
4 3 2 1 0 1 2 3 4
Gaussian distribution for linear distance
(a)
这些 危险 分子 严重 影响 正常 的 移民 政策
zhexie weixian fenzi yanzhong yingxiang zhengchang de zousi zhengce
2 1 0 1 2 3 4 5 6
(b)
影响 yingxiang 1
分子 fenzi 0
严重 yanzhong 2
政策 zhengce 2
这些 zhexie 1
危险 weixian 1
正常 zhengchang 3
移民 yimin 3
的 de 4
(c)

Figure 1: (a) NMT with the local attention. The black dotted box is the current source aligned word and the red dotted box is the predicted target word. (b) Linear distances for the source word "*fenzi*", for which the number denotes the linear distance. (c) Syntax-directed distances for source word "*fenzi*", for which the blue number represents syntax-directed distance between each word and "*fenzi*".

distance with the other source words. The decoder then focuses on the syntax-related source words within the syntax-distance constraint to compute a more effective context vector for predicting target word. Moreover, we further propose a *double context* NMT architecture, which consists of a global context vector and a syntax-directed local context vector from the global attention, to provide more translation performance for NMT from source representation. The experiments on the large-scale Chinese-to-English and English-to-German translation tasks show that the proposed approach achieves a substantial and significant improvement over the baseline system.

2 Background

2.1 Global Attention-based NMT

In NMT (Bahdanau, Cho, and Bengio 2015), the context of translation prediction relies heavily on *attention mechanism* and source input. Typically, the decoder computes a alignment score e_{ij} between each source annotation $\boldsymbol{h}_j$ and predicted target word y_i according to the previous decoder hidden state $\boldsymbol{s}_{i-1}$

$$e_{ij} = f(\boldsymbol{s}_{i-1}, \boldsymbol{h}_j), \tag{1}$$

where f is a RNN with GRU. Then all alignment scores are normalized to compute weight α_{ij} of each encoder state $\boldsymbol{h}_j$

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{J} exp(e_{ik})}. \tag{2}$$

Furthermore, the α_{ij} is used to weight all source annotations for computing current time-step context vector $\boldsymbol{c}_i^g$:

$$\boldsymbol{c}_i^g = \sum_{j=1}^{J} \alpha_{ij}\boldsymbol{h}_j. \tag{3}$$

Finally, the context vector $\boldsymbol{c}_i$ is used to predict target word y_i by a non-linear layer:

$$P(y_i|y_{<i}, x) = softmax(\boldsymbol{L}_o \mathbf{tanh}(\boldsymbol{L}_w \boldsymbol{E}_y[\hat{y}_{i-1}] + \boldsymbol{L}_d \mathbf{s}_i + \boldsymbol{L}_{cg}\mathbf{c}_i^g)) \tag{4}$$

where $\boldsymbol{s}_i$ is the current decoder hidden state and y_{i-1} is the previously emitted word; the matrices $\boldsymbol{L}_o$, $\boldsymbol{L}_w$, $\boldsymbol{L}_d$ and $\boldsymbol{L}_{cg}$ are transformation matrices. Intuitively, this attention is called as *global attention* because of the context vector $\mathbf{c}_i^g$ takes all source words into consideration (Luong, Pham, and Manning 2015).

2.2 Local Attention-based NMT

Compared with the *global attention*, the *local attention* selectively focuses on a small window of context (Luong, Pham, and Manning 2015). It first generates a source aligned position p_i for the predicted target word at current decoder time-step i:

$$p_i = J \cdot sigmoid(\boldsymbol{v}^T tanh(\boldsymbol{W}_p \boldsymbol{h}_i^{'})), \tag{5}$$

where J is the length of source sentence and $\boldsymbol{h}_i^{'}$ is decoder hidden state, $\boldsymbol{v}^T$ and $\boldsymbol{W}_p$ are weights.

To focus on source words within the fixed-window, the α_{ij}^l is refined by the follow eq.(6):

$$\alpha_{ij}^l = \begin{cases} \alpha_{ij} exp(-\frac{(j-p_i)^2}{2\sigma^2}), & j \in [p_i - D, p_i + D] \\ 0, & j \notin [p_i - D, p_i + D], \end{cases} \tag{6}$$

where [p_i-D, p_i+D] denotes the local window and the standard deviation is empirically set as $\sigma = \frac{D}{2}$.[1] Moreover, the local attention focuses on source annotations in window

[1] The D is set as *10* in local attention of (Luong, Pham, and Manning 2015).

$[p_i - D, p_i + D]$ to compute the current time-step local context vector $\boldsymbol{c}_i^l$:

$$\boldsymbol{c}_i^l = \sum_{j \in [p_i - D, p_i + D]} \alpha_{ij}^l \boldsymbol{h}_j. \tag{7}$$

Finally, the context vector $\boldsymbol{c}_i^l$ is then used to predict target word y_i by a non-linear layer:

$$P(y_i | y_{<i}, x) = softmax(\boldsymbol{L}_o \textbf{tanh}(\boldsymbol{L}_w \boldsymbol{E}_y[\hat{y}_{i-1}] + \boldsymbol{L}_d \mathbf{s}_i + \boldsymbol{L}_{cl} \boldsymbol{c}_i^l)), \tag{8}$$

where $\boldsymbol{s}_i$ is the current decoder hidden state and y_{i-1} is the previously emitted target word.

3 Syntax-Directed Attention

3.1 Syntax Distance Constraint

In NMT, the decoder computes the current context vector by weighting each encoder state with alignment weight to predict target word. Actually, these alignment weights are defined by the linear distance with the aligned source center position, such as the word "*fenzi*" in Figure 1(a). In other words, the greater the distance to the center position is, the smaller the contribution of the source word to the context vector is. Recently, the source long-distance dependency has been explicitly explored to enhance the encoder of NMT, thus improving target word prediction (Chen et al. 2017; Wu, Zhou, and Zhang 2017). This means that syntax context is beneficial for NMT. However, the existing NMT cannot adequately capture the source syntax context by the linear distance attention mechanism.

To address this issue, we propose a syntax distance constraint (**SDC**), in which we learn a SDC mask for each source word, as shown in Figure 2. Specifically, given a source sentence F with dependency tree T, each node denotes a source word x_j and the distance between two connected nodes is defined as *one*. We then traverse every word according to the order of source word, and compute the distances of all remaining words to the current traversed word x_j as its SDC mask m_j. Finally, we learn a sequence of SDC mask $\{m_0, m_1, ..., m_J\}$, and organize them as a $J * J$ matrix $\mathcal{M}$, in which J denotes the length of source sentence, and elements in each row denote the distances of all word to the row-index word,

$$\mathcal{M} = [[m_0], [m_1], ..., [m_J]]. \tag{9}$$

As shown in Figure 2, the third row denotes the syntax context mask of word "*fenzi*". Specifically, syntax distance of "*fenzi*" itself is zero; the syntax distances of "*zhexie*", "*weixian*", and "*yingxiang*" are one; the syntax distance of "*yanzhong*" and "*zhengce*" are two; the syntax distance of "*zhengchang*" and "*yimin*", and "*de*" are four, as shown the black dotted box in Figure 2.

3.2 Syntax-Directed Attention

To capture the source context with the SDC (in Section 3.1), we propose a novel syntax-directed attention (**SDAtt**) for NMT, as shown in Figure 3. The decoder first learn aligned

	这些 *zhexie*	危险 *weixian*	分子 *fenzi*	严重 *yanzhong*	影响 *yingxiang*	正常 *zhengchang*	的 *de*	移民 *yimin*	政策 *zhengce*
这些 *zhexie*	0	2	1	3	2	4	5	4	3
危险 *weixian*	2	0	1	3	2	4	5	4	3
分子 *fenzi*	1	1	0	2	1	3	5	3	2
严重 *yanzhong*	3	3	2	0	1	3	4	3	2
影响 *yingxiang*	2	2	1	1	0	2	3	2	1
正常 *zhengchang*	4	4	3	3	2	0	1	2	1
的 *de*	5	5	4	4	3	1	0	3	2
移民 *yimin*	4	4	3	3	2	2	3	0	1
政策 *zhengce*	3	3	2	2	1	1	2	1	0

Figure 2: Syntax distance constraint mask matrix $\mathcal{M}$ for the dependency-based Chinese sentence in Figure 1(c), in which each row denotes the syntax distance mask of one source word, for example the dotted black box is syntax distance constraint mask for source word "*fenzi*".

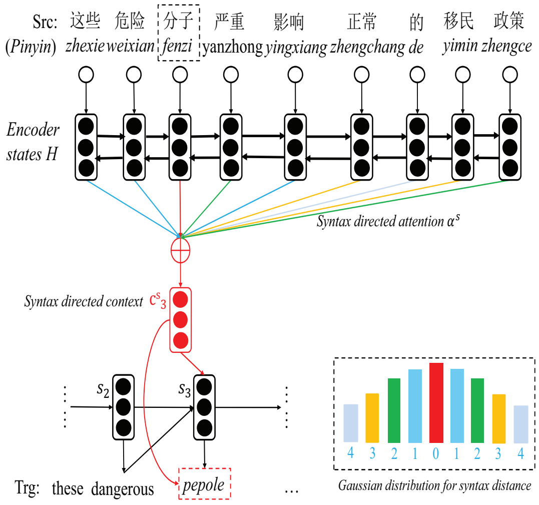

Src: 这些 危险 分子 严重 影响 正常 的 移民 政策
(Pinyin) zhexie weixian fenzi yanzhong yingxiang zhengchang de yimin zhengce
Encoder states H
Syntax directed attention α^s
Syntax directed context c^s_3
s_2
s_3
Trg: these dangerous pepole …
4 3 2 1 0 1 2 3 4
Gaussian distribution for syntax distance

Figure 3: Syntax-directed attention for NMT.

source position p_i of the current time-step i by the eq.(5). According to the position p_i, we obtain its SDC mask m_i from matrix $\mathcal{M}$ in eq.(9), i.e., $\mathcal{M}[p_i]$. We learn alignment score $e_i^s j$ with SDC mask $\mathcal{M}[p_i]$ by the following equation:

$$e_{ij}^s = e_{ij} exp(-\frac{(\mathcal{M}[p_i][j])^2}{2\sigma^2}), \tag{10}$$

where the Gaussian distribution centered around p_i is used to capture the difference of syntax distance, and further to tune the alignment score e_{ij} in eq.(1). Besides, the standard deviation σ is set as $\frac{n}{2}$, in which one syntax distance is different from one linear distance of the local attention. In other words, one syntax distance corresponds to multiple syntax-related words instead of two words in local attention. The n is more similar to the order of n-gram language model. Therefore, the n is empirically set as four in our ex-

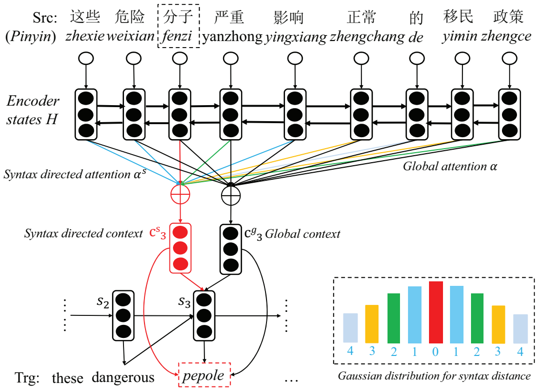

Src: 这些 危险 分子 严重 影响 正常 的 移民 政策
(Pinyin) zhexie weixian fenzi yanzhong yingxiang zhengchang de yimin zhengce
Encoder states H
Syntax directed attention α^s
Global attention α
Syntax directed context $c^s{}_3$
$c^g{}_3$ Global context
s_2
s_3
Trg: these dangerous pepole …
4 3 2 1 0 1 2 3 4
Gaussian distribution for syntax distance

Figure 4: Double context for NMT.

periments, which means that we only take *4*-gram SDC into account.

The $\alpha_{ij}^{s_n}$ is normalized within *n*-gram SDC:

$$\alpha_{ij}^{s_n} = \begin{cases} \frac{exp(e_{ij}^{s})}{\sum_{k \in M[p_i][k] \leq n} exp(e_{ik}^{s})}, & j \in [p_i - n, p_i + n] \\ 0, & j \notin [p_i - n, p_i + n], \end{cases} \tag{11}$$

In other words, we only consider words within the *n*-gram SDC and simply ignore the outside part of the *n*-gram SDC.

The context vector $\boldsymbol{c}_i^s$ is, then, computed as a weighted sum of these annotations $\boldsymbol{h}_i$ by alignment weights with the SDC:

$$\boldsymbol{c}_i^s = \sum_{j}^{J} \alpha_{ij}^{s_n} \boldsymbol{h}_j, \tag{12}$$

Finally, similar to the eq.(8), the context vector $\boldsymbol{c}_i^s$ is used to predict the target word y_i:

$$P(y_i|y_{<i}, x, T) = softmax(\boldsymbol{L}_o \mathbf{tanh}(\boldsymbol{L}_w \boldsymbol{E}_y[\hat{y}_{i-1}] + \boldsymbol{L}_d \boldsymbol{s}_i + \boldsymbol{L}_{cs} \boldsymbol{c}_i^s)) \tag{13}$$

where $\boldsymbol{s}_i$ is the current decoder hidden state and y_{i-1} is the previously emitted word.

3.3 Double Context Mechanism

In Section 3.2, the proposed SDAtt uses a local context with the SDC to compute current context vector instead of context vector with linear distance constraint in global or local attention. Inspired by the decoder with additional visual attention (Calixto, Liu, and Campbell 2017; Chen et al. 2017), we design a unique ***double context*** NMT as shown in Figure 4, to provide more translation performance for NMT from SDAtt in Section 3.2. The proposed model can be seen as an expansion of the global attention NMT framework described in Section 2.1 with the addition of a SDAtt to incorporate source syntax distance constraint.

Compared with the global attention, we learn two context vectors over a single global attention for target word prediction: a traditional (global) context vector which always attends to all source words and a syntax-directed context vector that focuses on *n*-gram (i.e., *4*-gram) source syntax context words. To that end, in addition to the traditional context vector $\boldsymbol{c}_i^g$ in eq.(3), we learn a context vector $\boldsymbol{c}_i^s$ for the SDC according to the eq.(12). Formally, the probability for the next target word is computed by the following eq.(14),

$$P(y_i|y_{<i}, x, T) = softmax(\boldsymbol{L}_o \mathbf{tanh}(\boldsymbol{L}_w \boldsymbol{E}_y[\hat{y}_{i-1}] + \boldsymbol{L}_d \boldsymbol{s}_i + \boldsymbol{L}_{cg} \boldsymbol{c}_i^g + \boldsymbol{L}_{cs} \boldsymbol{c}_i^s)). \tag{14}$$

4 Experiments

4.1 Data sets

The proposed methods were evaluated on two data sets.[2]

- For English (EN) to German (DE) translation task, *4.43* million bilingual sentence pairs of the WMT'14 data set was used as the training data, including Common Crawl, News Commentary and Europarl v7. The newstest2012 and newstest2013/2014/2015 was used as dev set and test sets, respectively.
- For Chinese (ZH) to English (EN) translation task, the training data set was *1.42* million bilingual sentence pairs from LDC corpora, which consisted of LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08, and LDC2005T06. The NIST02 and the NIST03/04/05/06/08 data sets were used as dev set and test sets, respectively.

4.2 Baseline Systems

Along with the standard phrase-based SMT (**PBSMT**) implemented in Moses (Koehn et al. 2007) and standard NMT with global attention (**GlobalAtt**) (Bahdanau, Cho, and Bengio 2015) baseline systems, we also compared the proposed methods to the recent related NMT methods:

- **Chen et al.(2017)**: extracted a local source dependency unit (including parent, siblings, and children of each source word) and learned its semantic representation. They introduced source dependency representation into the Encoder and Decoder by two kinds of NMT models, which extended source word with dependency representation and enhanced the global attention with dependency representation, respectively. Their methods are one of state-of-the-art syntax based NMT methods, which outperformed significantly the method of (Sennrich and Haddow 2016).
- **LocalAtt**: Luong et al. 2015 selectively computed alignment probabilities for fixed-window source words centering around current aligned source position instead of all source words.
- **FlexibleAtt**: Shu and Nakayama 2017 proposed a flexible attention NMT, which can dynamically create a window of the encoder states instead of fixed-window method of (Luong, Pham, and Manning 2015), and thus learned a flexible context to predict target word.
- **GlobalAtt+LocalAtt/FlexibleAtt**: we implemented the global attention with additional the local/flexible attention, to further evaluate our double context NMT.

ZH-EN	**Dev (NIST02)**	**NIST03**	**NIST04**	**NIST05**	**NIST06**	**NIST08**	**AVG**
PBSMT	33.15	31.02	33.78	30.33	29.62	23.53	29.66
GlobalAtt	37.12	35.24	37.49	34.60	32.48	26.32	33.23
Chen et al. (2017)	37.42	35.98	38.34	35.28	33.58	27.23	34.08
LocalAtt	37.31	35.57	37.85	34.93	32.74	26.83	33.58
FlexAtt	37.19	35.46	37.81	34.76	32.83	26.71	33.51
SDAtt	**38.01**	**36.67****†	**38.66****†	**35.74****†	**34.03****†	**27.66****†	**34.55**

EN-DE	**Dev (newstest2012)**	**newstest2013**	**newstest2014**	**newstest2015**	**AVG**
PBSMT	14.89	16.75	15.19	16.84	16.35
GlobalAtt	17.09	20.24	18.67	19.78	19.56
Chen et al. (2017)	17.48	21.03	19.43	20.56	20.31
LocalAtt	17.19	20.74	19.00	20.15	19.96
FlexibleAtt	17.24	20.57	19.12	20.03	19.91
SDAtt	**17.86**	**21.71****†	**20.36****†	**21.57****†	**21.21**

Table 1: Results on ZH-EN and EN-DE translation tasks for the proposed SDAtt. "*" indicates that the model significantly outperforms GlobalAtt at p-value$<$0.05, "**" indicates that the model significantly outperforms GlobalAtt at p-value$<$0.01. "†" indicates that the model significantly outperforms the best baseline Chen et al.2017's Model at p-value$<$0.05. **AVG** is the average BLEU score for all test sets. The bold indicates that the BLEU score of test set is better than the best baseline system.

ZH-EN	**Dev (NIST02)**	**NIST03**	**NIST04**	**NIST05**	**NIST06**	**NIST08**	**AVG**
PBSMT	33.15	31.02	33.78	30.33	29.62	23.53	29.66
GlobalAtt	37.12	35.24	37.49	34.60	32.48	26.32	33.23
+Chen et al. (2017)	38.11	37.35	39.00	36.12	33.78	27.81	34.81
+LocalAtt	37.89	37.06	38.73	36.10	33.62	27.43	34.59
+FlexibleAtt	37.97	36.86	38.56	35.62	33.94	27.37	34.47
+SDAtt	**38.61**	**38.19****†	**39.81****†	**36.74**$**$	**34.63****†	**28.61****†	**35.60**

EN-DE	**Dev (newstest2012)**	**newstest2013**	**newstest2014**	**newstest2015**	**AVG**
PBSMT	14.89	16.75	15.19	16.84	16.35
GlobalAtt	17.09	20.24	18.67	19.78	19.56
+Chen et al. (2017)	18.03	21.44	19.96	21.07	20.82
+LocalAtt	17.78	21.26	19.87	20.67	20.6
+FlexibleAtt	17.56	21.10	19.76	20.74	20.53
+SDAtt	**18.65**	**22.11****†	**20.75****†	**22.05****†	**21.64**

Table 2: Results on ZH-EN and EN-DE translation tasks for the double context mechanism.

All NMT models were implemented in the NMT toolkit Nematus (Sennrich et al. 2017).[3] We used the Stanford parser (Chang et al. 2009) to generate the dependency trees for source language sentences, such as Chinese sentences of ZH-EN and English sentences of EN-DE translation tasks. We limited the source and target vocabularies to *50*K, and the maximum sentence length was *80*. We shuffled training set before training and the mini-batch size is *80*. The word embedding dimension was *620*-dimensions and the hidden layer dimension was *1000*-dimensions, and the default dropout technique (Hinton et al. 2012) in Nematus was used on the all layers. Our NMT models were trained about *400*k mini-batches using ADADELTA optimizer (Zeiler 2012), taking six days on a single Tesla P100 GPU, and the beam size for decoding was *12*. Case-insensitive *4*-gram NIST BLEU score (Papineni et al. 2002) was as the evaluation metric, and the signtest (Collins, Koehn, and Kucerova 2005) was as statistical significance test.

4.3 Evaluating SDAtt NMT

Table 1 shows translation results on ZH-EN and EN-DE translation tasks for syntax-directed attention NMT in Section 3. The GlobalAtt significantly outperforms PBSMT by *3.57* BLEU points on average, indicating that it is a strong baseline NMT system. All the comparison methods, including Chen et al.(2017)'s model, LocalAtt, and FlexibleAtt, outperform the baseline GlobalAtt.

(1) Over the GlobalAtt, the proposed SDAtt gains an improvement of 1.32 BLEU points on average on ZH-EN translation task, which indicates that our method can effective improve translation performance of NMT.

(2) The SDAtt surpasses LocalAtt and FlexibleAtt by 0.97/1.04 BLEU points on average on ZH-EN translation task. This indicates that the proposed syntax distance constraint can capture more translation information to improve word prediction than linear distance constraint.

(3) The SDAtt also outperforms Chen et al.(2017)'s model on ZH-EN translation task by 0.47 BLEU points on average.

[2] Our method also was verified on the English-to-French translation task of the WMT'14 data set

[3] https://github.com/EdinburghNLP/nematus

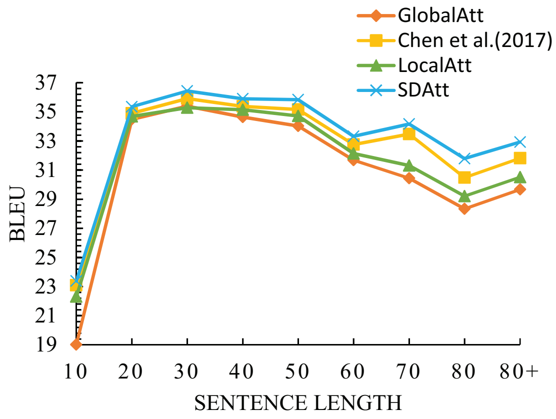


Figure 5: Translation qualities of different sentence lengths for SDAtt on the ZH-EN task.

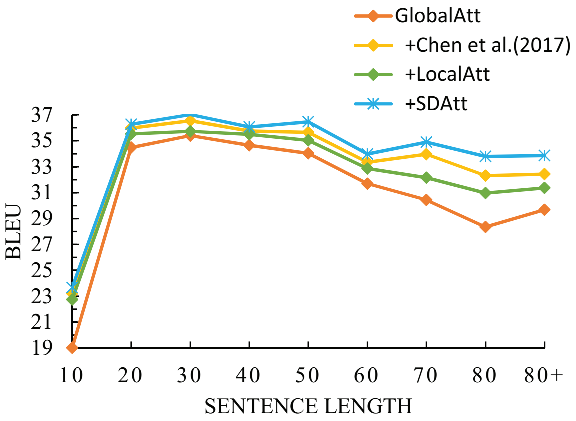


Figure 6: Translation qualities of different sentence lengths for GlobalAtt+SDAtt on the ZH-EN task.

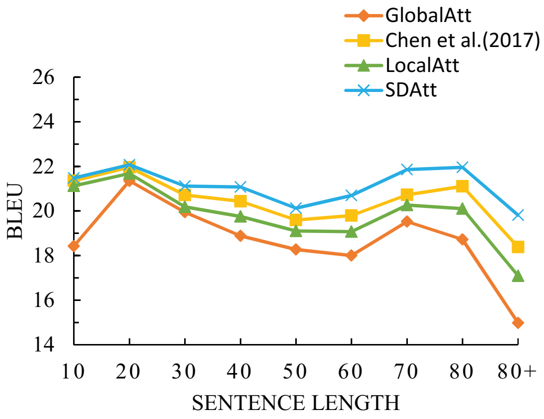


Figure 7: Translation qualities of different sentence lengths for SDAtt on the EN-DE task.

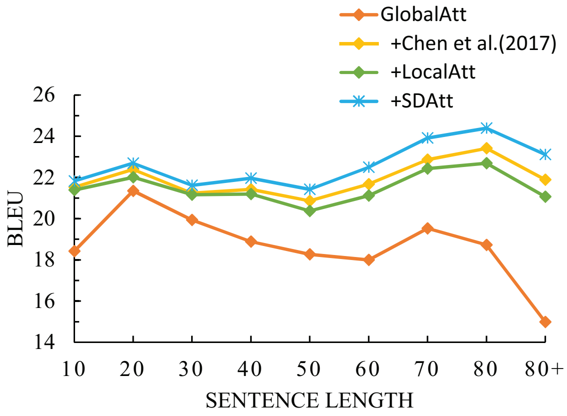


Figure 8: Translation qualities of different sentence lengths for GlobalAtt+SDAtt on the EN-DE task.

This shows that our method can capture more source dependency information to improve word prediction.

(4) For EN-DE translation task, the proposed SDAtt gives similar improvements over the baseline system and comparison methods. These results show that our method also can effectively improve the English-to-German translation task. In other words, the proposed SDAtt is a robust method for improving the translation of other language pairs.

4.4 Evaluating Double Context Mechanism

To further verify the effectiveness of the proposed double context mechanism, we compared it with three similar models, including +Chen et al. (2017)'s Model, +LocalAtt, and +FlexibleAtt. Table 2 showed translation results of the proposed double context method on ZH-EN and EN-DE translation tasks.

(1) All the comparison methods and our +SDAtt outperform the baseline GlobalAtt. In particularly, they gain further improvements by the corresponding single context NMT in Table 1, for example, +FlexibleAtt (34.81) *vs* LocalAtt (33.58). This indicates that the proposed double-context mechanism for NMT is more effective than single context NMT.

(2) The +SDAtt outperforms GlobalAtt by 2.37 BLEU points on average on ZH-EN translation task. Especially, the +SDAtt gains improvements of 1.01/1.13 BLEU points on average over the +LocalAtt/FlexibleAtt. This shows that the proposed SDAtt give more translation information for NMT from source representation.

(3) The +SDAtt outperforms +Chen et al.(2017)'s Model by 0.79 BLEU points on average on ZH-EN translation task. This means that the SDAtt is more effective than enhancing global attention with source dependency representation of Chen et al. (2017).

(4) For the EN-DE translation task, the proposed +SDAtt shows similar improvements over the baseline system and comparison methods. These results indicate that our double context architecture also can effectively improve the English-to-German translation task.

4.5 Effect of Translating Long Sentences

We grouped sentences of similar lengths on the test sets of the two tasks to evaluate the BLEU performance. For example, sentence length “*50*” indicates that the length of source sentences is between 40 and 50. We then computed a BLEU score per group, as shown in Figures 5-8.

Take ZH-EN task as a example in Figure 5 and 6, our methods, including SDAtt and +SDAtt, always yielded consistently higher BLEU scores than the baseline GlobalAtt in terms of different lengths. When the length came to “30”, they outperformed the best baseline Chen et al. (2017). This was because our methods can selectively focus on syntactic related source inputs with the current predicted target word and capture more source information to improve the performance of NMT. Moreover, our models also showed similar improvements for EN-DE task in Figures 7 and 8. This again showed the effectiveness of our method on long sentence translation.

5 Related Work

Recently, many efforts have been initiated on exploiting source- or target-side syntax information to improve the performance of NMT. Sennrich and Haddow (2016) augmented each source word with its corresponding part-of-speech tag, lemmatized form and dependency label. Li et al. (2017) linearized parse trees of source sentences to obtain structural label sequences, thus capturing syntax label information and hierarchical structures. To more closely combine the NMT with syntax tree, Eriguchi et al. (2017) proposed a hybrid model that learns to parse and translate by combining the recurrent neural network grammar into the attention-based NMT, and thus encouraged the NMT model to incorporate linguistic prior during training, and lets it translate on its own afterward. Wu et al. (2017) then proposed a sequence-to-dependency NMT model, which used two RNNs to jointly generate target translations and construct their syntactic dependency trees, and then used them as context to improve word generation. They extended source word with external syntax labels, thus providing richer context information for word prediction in NMT.

Eriguchi et al. (2016) proposed a tree-to-sequence attentional NMT, which use a tree-based encoder to compute the representation of the source sentence following its parse tree instead of the sequential encoder. It further was extended by bidirectional tree encoder which learns both sequential and tree structured representations (Huadong et al. 2017). Wu et al. (2017) enriched each encoder state from both child-to-head and head-to-child with global knowledge from the source dependency tree. Chen et al. (2017) extended each source word with local dependency unit to capture source long-distance dependency constraints, achieving an state-of-the-art performance in NMT, especially on long sentence translation. These methods focused on enhancing source representation by capturing syntax structures in the source sentence or target sentence, such as phrase structures and dependency structures for improving translation.

In this paper, we extend the local attention with a novel syntax distance constraint to capture syntax related encoder states with the predicted target word. Following the dependency tree of a source sentence, each source word has a distance mask, which denotes its syntax distances from the other source words. This mask is called as the syntax-distance constraint. The decoder then focuses on the syntax-related source words within this syntax-distance constraint to compute a more effective context vector for predicting target word. Moreover, we further propose a double context NMT architecture, which consists of a global context vector and a syntax-directed local context vector from the global attention, to provide more translation performance for NMT from source representation.

This work refines the local attention by syntax-distance constraint instead of traditional linear distance in the global or local attention, and thus selectively focuses on syntax-related source words to compute a more effective context vector for predicting target word.

6 Conclusion

In this paper, we explored the effect of syntactic distance on the attention mechanism. We then proposed a syntax-directed attention for NMT method to selectively focus on syntax related source words for predicting target word. Moreover, we further proposed a double context NMT architecture to provide more translation performance for NMT from source representation. In the future, we will exploit richer syntax information to improve the performance of NMT.

Acknowledgments

We thank Dr. Lemao Liu for invaluable discussions. We are grateful to the anonymous reviewers and the senior program member for their insightful comments and suggestions. This work is partially supported by the program “Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology” of MIC, Japan. Tiejun Zhao is supported by the National Natural Science Foundation of China (NSFC) via grant 91520204 and National High Technology Research & Development Program of China (863 program) via grant 2015AA015405.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of 6th International Conference on Learning Representations*.

Calixto, I.; Liu, Q.; and Campbell, N. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1913–1924. Vancouver, Canada: Association for Computational Linguistics.

Chang, P.-C.; Tseng, H.; Jurafsky, D.; and Manning, C. D. 2009. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*, 51–59. Boulder, Colorado: Association for Computational Linguistics.

Chen, K.; Wang, R.; Utiyama, M.; Liu, L.; and Akihiro Tamura, Eiichiro Sumita, T. Z. 2017. Neural machine translation with source dependency representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 23–32. Copenhagen, Denmark: Association for Computational Linguistics.

Cho, K.; van Merrienboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103–111. Doha, Qatar: Association for Computational Linguistics.

Collins, M.; Koehn, P.; and Kucerova, I. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 531–540. Ann Arbor, Michigan: Association for Computational Linguistics.

Eriguchi, A.; Hashimoto, K.; and Tsuruoka, Y. 2016. Tree-to-sequence attentional neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 823–833. Berlin, Germany: Association for Computational Linguistics.

Eriguchi, A.; Tsuruoka, Y.; and Cho, K. 2017. Learning to parse and translate improves neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada: Association for Computational Linguistics.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* abs/1207.0580.

Huadong, C.; Shujian, H.; David, C.; and Jiajun, C. 2017. Improved neural machine translation with a syntax-aware encoder and decoder. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada: Association for Computational Linguistics.

Kalchbrenner, N., and Blunsom, P. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709. Seattle, Washington, USA: Association for Computational Linguistics.

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A.; and Herbst, E. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 177–180. Prague, Czech Republic: Association for Computational Linguistics.

Li, J.; Deyi, X.; Zhaopeng, T.; Muhua, Z.; and Guodong, Z. 2017. Modeling source syntax for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada: Association for Computational Linguistics.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. Lisbon, Portugal: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Sennrich, R., and Haddow, B. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, 83–91. Berlin, Germany: Association for Computational Linguistics.

Sennrich, R.; Firat, O.; Cho, K.; Birch, A.; Haddow, B.; Hitschler, J.; Junczys-Dowmunt, M.; Läubli, S.; Miceli Barone, A. V.; Mokry, J.; and Nadejde, M. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 65–68. Valencia, Spain: Association for Computational Linguistics.

Shu, R., and Nakayama, H. 2017. An empirical study of adequate vision span for attention-based neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, 1–10. Vancouver: Association for Computational Linguistics.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 3104–3112. Cambridge, MA, USA: MIT Press.

Wu, S.; Zhang, D.; Yang, N.; Li, M.; and Zhou, M. 2017. Sequence-to-dependency neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 698–707. Vancouver, Canada: Association for Computational Linguistics.

Wu, S.; Zhou, M.; and Zhang, D. 2017. Improved neural machine translation with source syntax. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 4179–4185.

Zeiler, M. D. 2012. ADADELTA: an adaptive learning rate method. *CoRR* abs/1212.5701.