# Syntax-Directed Attention for Neural Machine Translation

**Kehai Chen**[1]*, **Rui Wang**[2]†, **Masao Utiyama**[2], **Eiichiro Sumita**[2], **and Tiejun Zhao**[1]
[1]Machine Intelligence and Translation Laboratory, Harbin Institute of Technology, Harbin, China
[2]ASTREC, 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-directed distance constraints. In this paper, we extend local attention with syntax-distance constraint, to focus on syntactically related source words with the predicted target word, thus learning a more effective context vector for word prediction. Moreover, we further propose a double context NMT architecture, which consists of a global context vector and a syntax-directed context vector over 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-Germen translation tasks show that the proposed approach achieves a substantial and significant improvement over the baseline system.

## 1 Introduction

Recent work of neural machine translation (**NMT**) have been proposed to adopt the encoder-decoder framework (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014), which employs a recurrent neural network (**RNN**) encoder to model source context information and a RNN decoder to generate translations Especially, Bahdanau et al. (2015) propose an NMT model with an *attention mechanism* (called as global attention), which acquires source sentence context dynamically at each decoding step, thus improving the performance of NMT. Luong et al. (2015) further refine the global attention into a local attention, 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 position $p_i$ for the current time-step $i$. A window of encoding states centered around the source position $p_i$ are weighted with alignment weights $\alpha$ to obtain a context vector $\mathbf{c}_i$. The weights $\alpha$ are inferred from the current target state and those encoder states in the window. Figure 1(a) shows a Chinese-to-English NMT model with local attention, and its contextual window is set to five. When the predicted 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 encoding states by Gaussian distribution, which penalty their alignment weights according to the distance with word "*fenzi*", such as their distances are {*2*, *1*, *0*, *1*, *2*} in turn. In other words, the greater the distance from the encoder states 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

*This paper has been accepted by the AAAI-2018. Kehai Chen was an internship research fellow at NICT when conducting this work.

†Corresponding author.

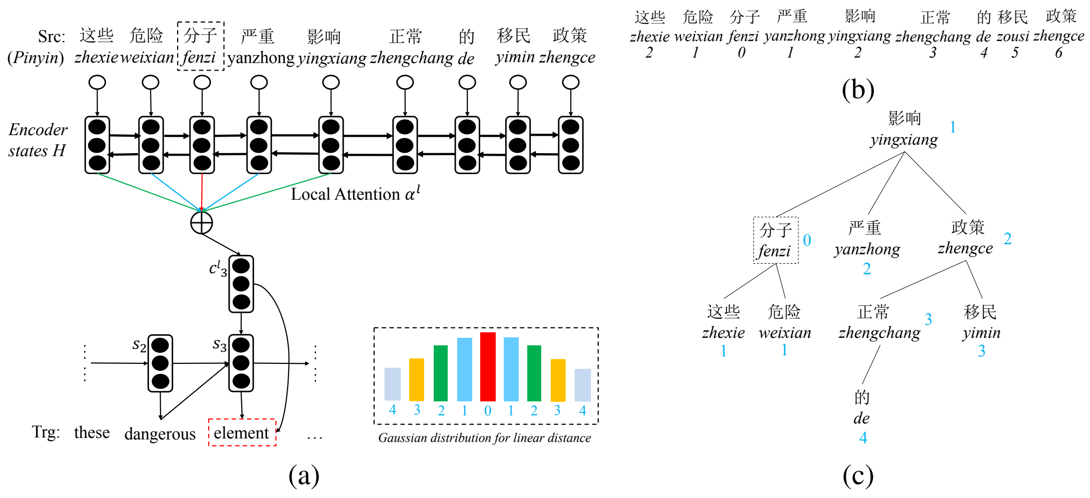


Figure 1: (a) NMT with local attention. The black dotted box is the current source aligned word and the red dotted box is the predicted target word. (b) Linear distances for the source word "*fenzi*", for which the number denotes the linear distance. (c) Syntax-directed distances for source word "*fenzi*",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

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 predicted central 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 on 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 global or local attention, does not allow NMT to focus on source context with the syntax distance constraint.

In this paper, we extend local attention with a novel syntax-distance constraint (**SDC**), to capture syntax related encoder states with the predicted target word. Following the dependency tree of a source sentence, each source word has a SDC mask, which denotes its syntax distance with the central source word. The decoder then focuses on the syntax-related source words within the SDC 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 global attention, to provide more translation performance for NMT from source representation. The experiments on the large-scale Chinese-to-English and English-to-Germen 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 et al., 2015), the context of translation prediction relies heavily on *attention mechanism* and source input. Typically, the decoder computes a alignment score $e_{ij}$ between source annotation $\mathbf{h}_j$ and predicted target word $y_i$ according to the previous decoder hidden state $\mathbf{s}_{i-1}$

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

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

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

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

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

Finally, the context vector $\mathbf{c}_i$ is then 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 $\mathbf{s}_i$ 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 (Luong et al., 2015) into consideration.

### 2.2 Local Attention-based NMT

Compared with the *global attention*, *local attention* selectively focuses on a small window of context (Luong et al., 2015). It first generates an aligned position $p_t$ for each target word at decode time-step $i$:

$$p_i = S \cdot sigmoid(\mathbf{v}^T tanh(\boldsymbol{W}_p \mathbf{h}_i)), \tag{5}$$

where $S$ is the length of source sentence and $\mathbf{h}_i$ is decoder hidden state, $\mathbf{v}^T$ and $\boldsymbol{W}_p$ are model parameters. The context vector $\mathbf{c}_i$ is then derived as a weighted summation over encoder states within the window [$p_i$-$D$, $p_i$+$D$], where $D$ is empirically selected[1]. Therefore, $\alpha_{ij}^l$ of each annotation $\mathbf{h}_j$ follows:

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

where the standard deviation is empirically set as $\sigma = \frac{D}{2}$. Moreover, the local attention focuses on source annotations in window $[p_i - D, p_i + D]$ to compute the current time-step local context vector $\mathbf{c}_i^l$:

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

[1]The $D$ is set as 10 in local attention of (Luong et al., 2015)

Finally, the context vector $c_i^l$ is then used to predict target word $y_i$ by a non-linear layer:

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

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

## 3 Syntax-Directed Attention

### 3.1 Syntax Distance Constraint

In NMT, the decoder computes current context vector by weighting each encoder state with alignment weight to predict target word. Actually, these alignment weights are 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. Recent works of (Chen et al., 2017) and Wu et al. 2017b explicitly encoded source long-distance dependency context to improve target word prediction. 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.

| | 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: SDC 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 source context mask for source word "*fenzi*".

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

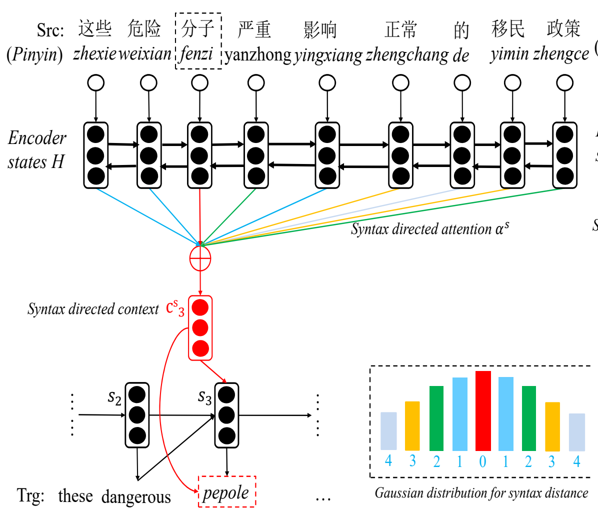


Figure 3: Syntax-distance attention for NMT.

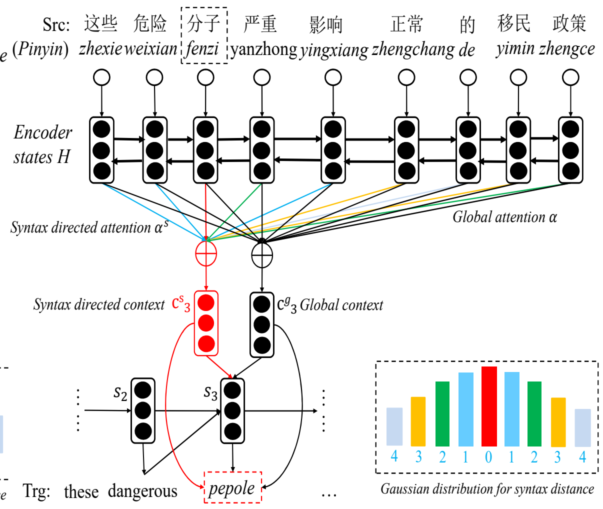


Figure 4: Double attention for NMT.

Figure 2. Specifically, given a source sentence $F$ with dependency tree $T$, each node denotes a source word $x_i$ 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 $x_i$ as its syntax context mask $m_i$. Finally, we learn a sequence of SDC mask $\{m_0, m_1, ..., m_L\}$, and organize them as a $S * S$ matrix $\mathcal{M}$, in which $S$ denotes the length of source sentence, and elements in each row denote the distances of all word to the row-index word.

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

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

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

where the Gaussian distribution centered around $p_i$ is used to capture the difference of SDD, and further to tune alignment score $e^g_{ij}$ in eq.(1). Besides, the standard deviation $\sigma$ is set as $\frac{n}{2}$, in which the $n$ is different from the $D$ of the local attention. In other words, one SDC corresponds to multiple syntax context word instead of two words in local attention. The $n$ is more similar to the order of language model. Therefore, the $n$ is empirically set as four in our experiments, which means that we only take into account 4-gram SDC.

When we only take into account $n$-gram (i.e., *4*-gram) SDC, the $\alpha^{s_n}_{ij}$ is normalized within $n$-gram SDC:

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

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

The context vector $\mathbf{c}^s_i$ is, then, computed as a weighted sum of these annotations $h_i$ by alignment weights with the SDC:

$$\mathbf{c}^s_i = \sum_j^J \alpha^{s_n}_{ij} \mathbf{h}_j, \tag{12}$$

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

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

where $\mathbf{s}_i$ 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 et al., 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-based NMT framework described in Section 2.1 with the addition of a SDAtt to incorporate source syntax distance constraint.

Compared with global attention-based NMT, 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 $\mathbf{c}_i^g$ in eq.(3), we learn a context vector $\mathbf{c}_i^s$ for the SDC according to the eq.(12). Formally, the probability for the next target word is computed by the following eq.(14)

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

## 4 Experiments

### 4.1 Data sets

The proposed methods were evaluated on two data sets.

- 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/newstest2014/newstest2015 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/NIST04/NIST05/NIST06/NIST08 data sets were used as dev set and test sets, respectively.

### 4.2 Baseline Systems

Along with standard phrase-based SMT (**PBSMT**) implemented in Moses (Koehn et al., 2007) and standard attentional NMT (**Global Attention**) Bahdanau et al.(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 (Chen et al., 2017) and Chen et al. (Chen et al., 2017)+Global Attention), respectively. Their method is one of state-of-the-art syntax based NMT methods, which outperformed Sennerich and Haddow (2016)' method significantly.
- **LocalAtt**: Luong et al. (2015) selectively compute 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 in Luong et al. (2015), and thus learned a flexible context to predict target word.
- **GlobalAtt+LocalAtt/FlexibleAtt**: we implemented global attention with additional local/flexible attention, to further evaluate our double context based NMT.

All NMT models are implemented in NMT toolkit Nematus (Sennrich et al., 2017). We use the Stanford parser (Chang et al., 2009) to

| 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 |
| Global Attention | 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 |
| Local Attention | 37.31 | 35.57 | 37.85 | 34.93 | 32.74 | 26.83 | 33.58 |
| Flexible Attention | 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 Global Attention at a $p$-value<0.05, "**" indicates that the model significantly outperforms Global Attention at a $p$-value<0.01. "†" indicates that the model significantly outperforms the best baseline Chen et al.2017's Model at a $p$-value<0.05. **AVG** is the average BLEU score for all test sets. The Bold indicates that the BlUE 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.

generate dependency trees for source language sentences, such as Chinese sentences of ZH-EN and English sentences of EN-DE translation tasks. We limit the source and target vocabularies to *50*K, and the maximum sentence length is *80*. We shuffle training set before training and the mini-batch size is *80*. The word embedding dimension is *620*-dimensions and the hidden layer dimension is *1000*, and the default dropout technique (Hinton et al., 2012) in Nematus is used on the all the layers. Our NMT models are trained about *400*k mini-batches using ADADELTA optimizer (Zeiler, 2012), taking 6 days on a single Tesla P100 GPU, and the beam size for decoding is *12*. Case-insensitive *4*-gram NIST BLEU score (Papineni et al., 2002) is as evaluation metric, and the signtest (Collins et al., 2005) is as statistical significance test.

### 4.3 Evaluating SDAtt NMT

Table 1 shows translation results on ZH-EN and EN-DE translation tasks for syntax-based attention NMT in Section 3. The Global Attention 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, Local Attention, and Flexible Attention, outperform the baseline Global Attention.

(1) Over the Global Attention, Syntax-distance Attention 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) Syntax-distance Attention surpasses Local Attention and Flexible Attention by 0.97/1.04 BLEU points on average on ZH-EN translation task. This indicate that the proposed syntax-

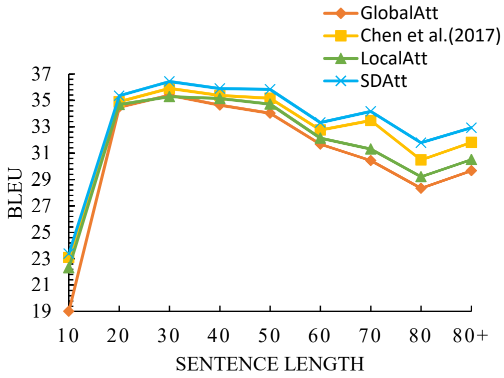


Figure 5: Translation qualities of different sentence lengths for SDAtt on the ZH-EN task.

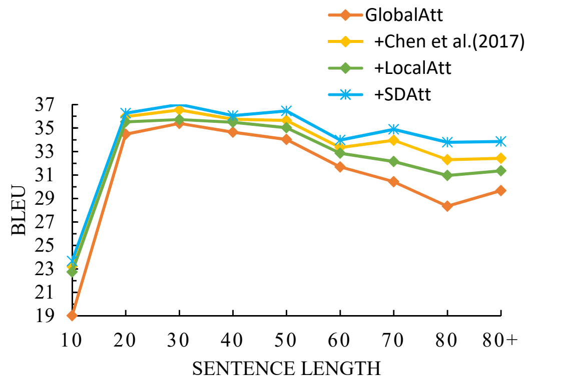


Figure 6: Translation qualities of different sentence lengths for GlobalAtt+SDAtt on the ZH-EN task.

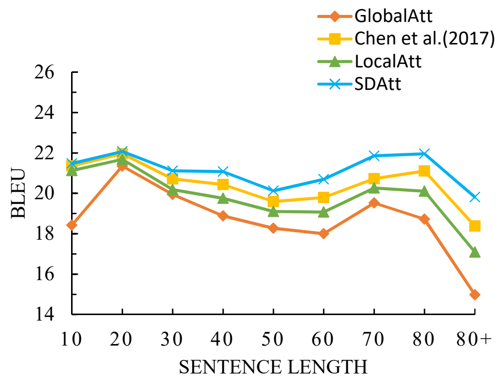


Figure 7: Translation qualities of different sentence lengths for SDAtt on the EN-DE task.

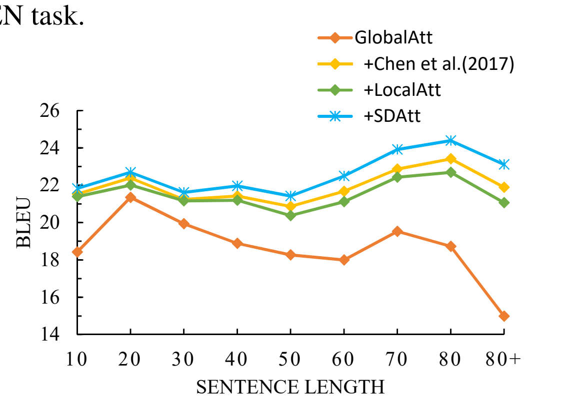


Figure 8: Translation qualities of different sentence lengths for GlobalAtt+SDAtt on the EN-DE task.

distance attention can capture more translation information to improve word prediction than linear distance attention.

(3) Syntax-distance Attention also outperforms Chen et al. (2017) 's model on ZH-EN translation task by 0.47 BLEU points on average. This shows that our method can learn more source dependency information to improve word prediction.

(4) For EN-DE translation task, the proposed Syntax Attention 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 architecture, we compare it with three similar models, including +Chen et al. (2017)'s Model, +LocalAtt, and +FlexibleAtt. Table 2 shows translation results of the proposed double attention 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 architecture for NMT is more effective than single context architecture for NMT.

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

(3) +SDAtt outperforms +Chen et al. (2017) by 0.79 BLEU points on average on ZH-EN translation task. This means that the syntax-

directed context is better than extending each source words with local dependency unit of Chen et al. (2017)'s Model.

(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 group sentences of similar lengths on the test sets of the two tasks to evaluate the BLEU performance. For example, sentence length "*40*" indicates that the length of source sentences is between 30 and 40. We then compute a BLEU score per group, as shown in Figures 5-8.

Take ZH-EN task as a example in Figure 5 and 6, clearly, our methods, including SDAtt and +SDAtt, always yield consistently higher BLEU scores than the baseline GlobalAtt in terms of different lengths. When the length comes to "30", they outperform the best baseline Chen et al. (2017)'s methods. This is because our methods can selectively focus on syntactic related source inputs with the current predicted target word and provide more effective source information to improve the performance of NMT. Moreover, our models also show similar improvements for EN-DE task in Figures 7 and 8. This again shows 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. (2017a) 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. (2017b) 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 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 vector, which denotes its syntax distance with the central source word. This mask is called as syntax-distance constraint. The decoder then focuses on the syntax-related source words within this syntax-distance constraint to compute an more effective context vector for predicting target word. Moreover, we further propose a double attention NMT architecture, which consists of a global context vector (from global attention) and a syntax-based local context vector (only from syntax-based local attention), to relieve more translation performance for NMT from source representation.

This work refines the local attention by syntax-distance constraint instead of traditional linear distance in 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 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 would like to thank the anonymous reviewers and the senior program member for their insightful comments and suggestions. This work was 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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of 6th International Conference on Learning Representations*.

Iacer Calixto, Qun Liu, and Nick Campbell. 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)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*, pages 51–59, Boulder, Colorado. Association for Computational Linguistics.

Kehai Chen, Rui Wang, Masao Utiyama, Lemao Liu, and Tiejun Zhao Akihiro Tamura, Eiichiro Sumita. 2017. Neural machine translation with source dependency representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 23–32, Copenhagen, Denmark. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 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*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 531–540, Ann Arbor, Michigan. Association for Computational Linguistics.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-sequence attentional neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 823–833, Berlin, Germany. Association for Computational Linguistics.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 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.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Chen Huadong, Huang Shujian, Chiang David, and Chen Jiajun. 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.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Junhui Li, Xiong Deyi, Tu Zhaopeng, Zhu Muhua, and Zhou Guodong. 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.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 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*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

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

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

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

Shuangzhi Wu, Dongdong Zhang, Nan Yang, Mu Li, and Ming Zhou. 2017a. Sequence-to-dependency neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–707, Vancouver, Canada. Association for Computational Linguistics.

Shuangzhi Wu, Ming Zhou, and Dongdong Zhang. 2017b. Improved neural machine translation with source syntax. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4179–4185.

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