IEICE TRANS. INF. & SYST., VOL.E103–D, NO.3 MARCH 2020

PAPER

Neural Machine Translation with Target-Attention Model

Mingming YANG[†a)], ***Nonmember***, **Min ZHANG**[†,††], ***Member***, **Kehai CHEN**[†††], **Rui WANG**[†††], ***and*** **Tiejun ZHAO**[†], ***Nonmembers***

SUMMARY *Attention mechanism*, which selectively focuses on source-side information to learn a context vector for generating target words, has been shown to be an effective method for neural machine translation (**NMT**). In fact, generating target words depends on not only the source-side information but also the target-side information. Although the vanilla NMT can acquire target-side information implicitly by recurrent neural networks (**RNN**), RNN cannot adequately capture the global relationship between target-side words. To solve this problem, this paper proposes a novel target-attention approach to capture this information, thus enhancing target word predictions in NMT. Specifically, we propose three variants of target-attention model to directly obtain the global relationship among target words: 1) a *forward target-attention model* that uses a target attention mechanism to incorporate previous historical target words into the prediction of the current target word; 2) a *reverse target-attention model* that adopts a reverse RNN model to obtain the entire reverse target words information, and then to combine with source context information to generate target sequence; 3) a *bidirectional target-attention model* that combines the forward target-attention model and reverse target-attention model together, which can make full use of target words to further improve the performance of NMT. Our methods can be integrated into both RNN based NMT and self-attention based NMT, and help NMT get global target-side information to improve translation performance. Experiments on the NIST Chinese-to-English and the WMT English-to-German translation tasks show that the proposed models achieve significant improvements over state-of-the-art baselines.

key words: *attention mechanism, neural machine translation, forward target-attention model, reverse target-attention model, bidirectional target-attention model*

1. Introduction

Recent works of neural machine translation (**NMT**) have been proposed to adopt the encoder-decoder framework [1], which employs a recurrent neural network (**RNN**) encoder to represent a source sentence as a sequence of vectors, which is fed into an RNN decoder to generate target translation word by word. Especially, the NMT with an *attention mechanism* is proposed to acquire a context vector over a sequence of vectors dynamically at each decoding step, thus improving the performance of NMT [2]. In NMT attention models, RNN-based [2], CNN-based [3], and self-attention-based [4] are imported. Many studies [2]–[5] have shown that attention mechanism is able to effectively detect the dependency relationship between all source inputs and the next predicted target word at each decoding step. However, the vanilla attention NMT focuses on source-side information to learn a dependent-time context vector for generating target word by the attention mechanism and ignores target-side global dependencies between the current predicted target word and the other target words, including the previous and the future target-side words.

Table 1 shows a Chinese-to-English translation example of NMT. The Chinese word "多少" has two kinds of meaning. One is "rather", the other is "how many". We observe that the Chinese word "多少" is not translated into "rather" due to the failure of capturing enough information from the forward target-side word "way" and the backward target-side word "pity". The neglect of these important clues may be due to the inefficiency of capturing global target-side relationship using the decoder hidden state learned by RNN or self-attention*. However, the target-side information may be beneficial for improving target word translation in NMT since they provide global relationship information among target words. In this paper, we propose a simple yet effective target-attention approach to take advantage of the entire target-side context information in the NMT system explicitly. To this end, we propose three kinds of NMT models for the target-attention:

- *Forward target-attention model:* An additional target-attention is learned based on all of the historical hidden states to gain a forward target context vector, and thus predict translation together with the existing source context vector.
- *Reverse target-attention model:* In contrast to the forward target-attention model, the reverse attention model is learned over the reversing target-side words

Table 1 An example of Chinese-to-English translation. The translation of the Chinese words in red needs forward and backward sentence information of the English sentence.

Src	关键 是 比赛 过程 ， **多少** 令人 感到 失望 。
Ref	the key problem is that the **way** went make people feel it was **rather** a **pity** .
NMT	the key is the competition process, **how many** people feel regret .

*Self-attention can only acquire the previous target information and ignore the future target information.

Manuscript received June 7, 2019.
Manuscript revised September 18, 2019.
Manuscript publicized November 26, 2019.

†The authors are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China.
††The author is with the School of Computer Science and Technology, Soochow University, Suzhou, 215006, China.
†††The authors are with National Institute of Information and Communications Technology, Kyoto-shi, 619–0289 Japan.
a) E-mail: mmyang@hit-mtlab.net (Corresponding author)

DOI: 10.1587/transinf.2019EDP7157

for capturing reverse relationship among target-side context.

- *Bidirectional target-attention model:* To further improve translation performance from target-side information, both of the forward and reverse target-attentions are integrated into the vanilla NMT to predict translations.

2. Attention-Based NMT

In this section, we introduce the background of the RNN based NMT [2] and the Transformer based NMT [4].

2.1 RNN Based NMT

In the RNN based NMT, the encoder applies bidirectional recurrent neural networks (**Bi-RNN**) to encode a source sentence: one reads an input sequence $X = (x_1, x_2, \ldots, x_J)$ from left to right and outputs a forward sequence of hidden states sequence $(\overrightarrow{h}_1, \overrightarrow{h}_2, \cdots, \overrightarrow{h}_J)$, $\overrightarrow{h}_j = \overrightarrow{RNN}(x_j, \overrightarrow{h}_{j-1})$. While the other operates from right to left and outputs a backward hidden states sequence $(\overleftarrow{h}_1, \overleftarrow{h}_2, \cdots, \overleftarrow{h}_J)$, $\overleftarrow{h}_j = \overleftarrow{RNN}(x_j, \overleftarrow{h}_{j+1})$. Where $\overrightarrow{RNN}$ or $\overleftarrow{RNN}$ are a RNN with GRU or LSTM, our work is based on RNN with GRU which is smaller and faster than LSTM. The final annotation vector is the concatenation of forward and backward vectors: $h_j = [\overrightarrow{h}_j; \overleftarrow{h}_j]$. The encoder represents source input sentence as a sequence of source annotation vectors $H = (h_1, h_2, \cdots, h_J)$. The decoder is also a RNN that predicts a target sequence $Y = (y_1, y_2, \ldots, y_I)$. The hidden state s_i of decoder at time step i is computed:

$$s_i = f(s_{i-1}, y_{i-1}, c_i), \tag{1}$$

where $f(\cdot)$ is GRU unit, a highly non-linear function. The implementation is shown below:

$$\begin{aligned} r_i &= \sigma(W_r y_{i-1} + U_r s_{i-1} + V_r c_i + b_r), \\ u_i &= \sigma(W_u y_{i-1} + U_u s_{i-1} + V_u c_i + b_u), \\ \hat{s}_i &= \tanh(W y_{i-1} + U[r_i \odot s_{i-1}] + V c_i + b), \\ s_i &= 1 - u_i \odot s_{i-1} + u_i \odot \hat{s}_i, \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes the element-wise multiplication. W_r, W_u, W, U_r, U_u, U, V_r, V_u, V, b_r, b_u, b are the parameters of the model, r_i and u_i are update and reset gates of GRU, respectively.

In the attention model, the current context vector c_i is calculated as a weighted sum over source annotation vectors $(h_1, h_2, \cdots, h_J)$ with alignment weights $\alpha_{i,j}$:

$$c_i = \sum_{j=1}^{J} \alpha_{i,j} h_j, \tag{3}$$

where $\alpha_{i,j}$ is the scalar weight of each hidden state $\boldsymbol{h}_j$ computed by the attention model and a is a feedforward neural network:

$$\begin{aligned} \alpha_{i,j} &= \frac{\exp(e_{i,j})}{\sum_{j'=1}^{J} \exp(e_{i,j'})}, \\ e_{i,j} &= a(s_{i-1}, h_j). \end{aligned} \tag{4}$$

The translation probabilities of next target word y_i are computed via multi-layer perception neural network g, which is based on the current decoder hidden state s_i, the previous word y_{i-1} and a current source-side context vector c_i:

$$P(y_i | y_{<i}; X) = g(y_{i-1}, s_i, c_i). \tag{5}$$

2.2 Transformer Based NMT

Transformer [4] is also an encoder-to-decoder architecture. Different from the other NMT, it has the self-attention layers **(SAN)** that can operate in parallel. Each single self-attention layer has two sublayers: a multi-head self-attention layer and a feed forward network. Both sublayers are stacked using residual connection and layer normalization. Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions, which is formulated as follows:

$$\begin{aligned} &\text{MultiHead}(Q, K, V) = \text{Concat}_{i=1}^{s} h_i(q, k, v), \\ &h_i(q, k, v) = \text{Attention}(\frac{qW_i^q}{\sqrt{d_s}}, kW_i^k, vW_i^v), \end{aligned} \tag{6}$$

each head uses parameter matrices W_i^q, W_i^k and $W_i^v \in \mathbb{R}^{d \times d_s}$ to transform the input q, k, v, where d_s is a scale factor, which equals to d/s, d is the hidden size of q, and s is the number of heads.

The feed forward network consists of two linear transformations with a ReLU activation in between:

$$\text{FeedForwad}(x) = f_2(\text{Max}(0, f_1(x))), \tag{7}$$

where f_1 and f_2 are both feedforward networks. For the sake of brevity, we refer the reader to Vaswani *et al.* [4] for more details.

Denote H_{enc} as the representation of source sentences via the SAN of the encoder, and F_{dec} is also the representation of decoder by the SAN, Which can be computed as follows:

$$\begin{aligned} H_{enc} &= \text{Attention}(Q_x, K_x, V_x), \\ H_{dec} &= \text{Attention}(Q_y, K_y, V_y), \\ F_{dec} &= \text{Attention}(H_{dec}, H_{enc}, H_{enc}), \end{aligned} \tag{8}$$

where $Q_x = K_x = V_x$ are a source input sequence X, and $Q_y = K_y = V_y$ are a target predict sequence Y. The parameters of Transformer are trained to minimize the following objective function on a set of training examples $\{(X^n, Y^n)\}_{n=1}^{N}$:

$$L(\theta) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{I_y} \log P(y_i^n | y_{<i}^n, H_{enc}, F_{dec}; \theta), \tag{9}$$

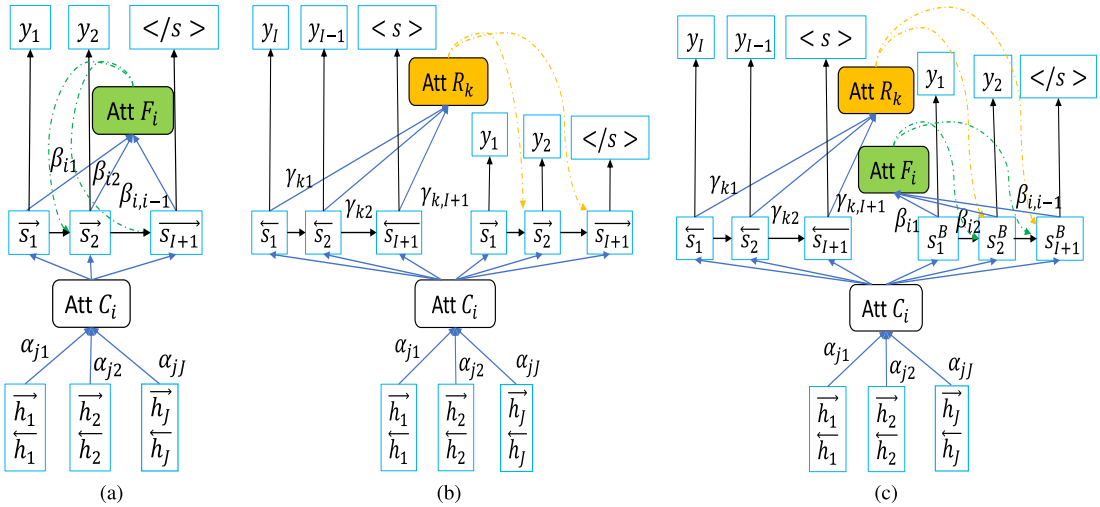

Fig. 1 (a) Architecture of RNN with forward target-attention model; (b) architecture of RNN with reverse target-attention model; (c) architecture of RNN with bidirectional target-attention model.

where θ is a set of model parameters and $y_{<i}$ denotes a partial translation.

3. NMT with Target-Attention

Different from the conventional attention-based NMT which generates current target word with the previous decoder hidden state, all previous historical hidden states are taken into account in our target-attention models. To take full advantage of target-side information, we propose three kinds of target-attention models: 1) *Forward target-attention model*; 2) *Reverse target-attention model*; 3) *Bidirectional target-attention model*.

3.1 Forward Target-Attention Model

Figure 1 (a) illustrates our forward target-attention model. In this model, the encoder is the same as that of the traditional NMT. Compared with traditional NMT, the forward target-attention model aims to explore all previous historical decoder hidden states for predicting target word instead of an only single previous decoder hidden state. We consider that the target-side information can help NMT improve target word translation since it can capture additional long-distance relationship among target-side historical words. To this end, an dynamic list, which stores all previous target historical hidden states $D^f_{i-1} = (\overrightarrow{S}_1, \overrightarrow{S}_2, \cdots, \overrightarrow{S}_{i-1})$ is firstly added into the decoder of NMT. When generating the current target word y_i, we then compute a forward target-attention F_{i-1} with the dynamic list D^f as:

$$\begin{aligned} \beta_{i,i'} &= \frac{\exp(d_{i,i'})}{\sum_{i'=1}^{i-1}\exp(d_{i,i'})}, \\ d_{i,i'} &= b(\overrightarrow{S}_{i-1}, \overrightarrow{S}_{i'}), \end{aligned} \tag{10}$$

where b is a single feedforward neural network, and $\beta_{i,i'}$ is a normalized weight of each target historical hidden state $\overrightarrow{S}_{i'}$ computed by the forward target attention model.

The current target-side forward context vector F_{i-1} is calculated as a weighted sum over target historical hidden states in the dynamic list D^f with alignment weights $\beta_{i,i'}$:

$$F_{i-1} = \sum_{i'=1}^{i-1}\beta_{i,i'}\overrightarrow{S}_{i'}. \tag{11}$$

Finally, the learned F_{i-1} is as an additional input of the Eq. (1) to compute the current decoder hidden state $\overrightarrow{S}_i$:

$$\overrightarrow{S}_i = f(\overrightarrow{S}_{i-1}, y_{i-1}, c_i, F_{i-1}), \tag{12}$$

where $f(\cdot)$ is GRU unit, similar to Eq. (2). Meanwhile, the F_{i-1} is integrated into the computation of the conditional probability of the next word y_i:

$$P(y_i|y_{<i}; X) = g(y_{i-1}, \overrightarrow{S}_i, c_i, F_{i-1}). \tag{13}$$

We train the proposed NMT with forward target-attention a set of train data $\{(X^n, Y^n)\}_{n=1}^N$. Finally, there is an available NMT model with forward target-attention parameterized by θ_1, the objective is to minimize the following conditional probability:

$$L(\theta_1) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{I_y}\log P(y_i^n|y_{<i}^n, X^n; \theta_1). \tag{14}$$

We aim to make full use of the historical target-side information, so we set the dynamic list to store the forward target-side context information and a matrix of the attention mechanism which can learn the combined weights of the forward information. In the back propagation weight training, the matrix is only updated not the dynamic list.

3.2 Reverse Target-Attention Model

In the traditional n-gram language model, there is a strong

connection between the current word and the succeeding words [6]. In other words, the succeeding words are also beneficial for machine translation. However, these future relationships that not considered in the target-side of the NMT model.

In the vanilla NMT, there is a fact that source representations H, which encode not only forward source input sentence but also backward input sequence by BiRNN, is used to generate forward target language sequence. In other words, source representations H can also be used to generate a backward target language sequence. Therefore, to capture target-side future relationship, we add an additional RNN to obtain a reverse target-side hidden state $\overleftarrow{S}_k$ at each time-step k. A dynamic list D_k^r, which is similar to D_i^f in forward target attention model, for these learned reverse target-side historical hidden states $(\overleftarrow{S}_1, \overleftarrow{S}_2, \cdots, \overleftarrow{S}_k)$. Formally, the above procedure is similar to a decoder of the attention-based NMT:

$$P(\overleftarrow{y}_k|\overleftarrow{y}_{<k}; X) = g(\overleftarrow{y}_{k-1}, \overleftarrow{S}_{k-1}, c_k), \tag{15}$$

the difference is that the generated translation is a reverse target language sequence.

At each time-step k, we compute an alignment weight $\gamma_{k,k'}$ for each reverse historical target-side hidden state as follows:

$$\begin{aligned} \gamma_{k,k'} &= \frac{\exp(m_{k,k'})}{\sum_{k'=1}^{k-1}\exp(m_{k,k'})}, \\ m_{k,k'} &= q(\overleftarrow{S}_{k-1}, \overleftarrow{S}_{k'}), \end{aligned} \tag{16}$$

where q is also a single feedforward neural network.

According to the Eq. (3), the reverse target context vector R_k is calculated as a weighted sum over reverse target-side historical hidden states in the dynamic list D_k^r with alignment weights $\gamma_{k,k'}$:

$$R_k = \sum_{k'=1}^{k-1} \gamma_{k,k'} \overleftarrow{S}_{k'}, \tag{17}$$

The learned R_k is as an additional input of the Eq. (1) to compute the current decoder hidden state $\overleftarrow{S}_k$:

$$\overleftarrow{S}_k = f(\overleftarrow{S}_{k-1}, \overleftarrow{y}_{k-1}, c_k, R_{k-1}), \tag{18}$$

where $f(\cdot)$ is GRU unit, similar to Eq. (2). In order to make full use of all future target-side information and solve the problem that the length of reverse sequence and forward sequence may be inconsistent in the inference, we use the average of all reverse hidden states $\overleftarrow{S}$ as reverse future representations $\widetilde{R}$. Some studies showed that the **average** operation is an effective method to represent sentence [7]–[9], especially for NMT [10]. Compared to the traditional NMT, we add the reverse target context vector $\widetilde{R}$ into the conditional probability formula as follows:

$$P(y_i|y_{<i}; X) = g(y_{i-1}, \overrightarrow{S}_i, c_i, \widetilde{R}). \tag{19}$$

Due to the attention is based on reverse target-side historical hidden states, we call it a reverse target-attention model as shown in Fig. 1 (b).

To ensure the correctness of the target-side historical hidden states, we train both the source-to-forward_target translation model with reverse translation and the source-to-reverse_target translation model on a set of training examples $\{(X^n, Y^n)\}_{n=1}^N$:

$$\begin{aligned} L(\theta_2) = -\frac{1}{N}\Big\{ &\sum_{n=1}^{N}\sum_{i=1}^{I_y} \log P(y_i^n|y_{<i}^n, X^n; \theta_2) \\ &+ \sum_{n=1}^{N}\sum_{i=1}^{I_y} \log \overleftarrow{P}(\overleftarrow{y}_i^n|\overleftarrow{y}_{>i}^n, X^n; \theta_2)\Big\}. \end{aligned} \tag{20}$$

Finally, there is an available NMT model with reverse target attention parameterized by θ_2.

3.3 Bidirectional Target-Attention Model

RNN: although the previous two models have clearly employed the forward and the reverse semantic information between the target words, the current target-side word depends on both directional information. Therefore, we further propose a target-side bidirectional attention model to unite the forward and the reverse target-attention. Specifically, both of forward target context vector F_i in Eq. (11) and reverse target context vector $\widetilde{R}$ in (19) are used to compute the current decoder hidden state S_i^B as follows:

$$S_i^B = f(S_{i-1}^B, y_{i-1}, c_i, F_{i-1}, \widetilde{R}), \tag{21}$$

where $f(\cdot)$ is is the same as introduced in Eq. (2). Finally, our the conditional probability $p(y_i|y_{<i}; X)$ is formulated in Eq. (22):

$$P(y_i|y_{<i}; X) = g(y_{i-1}, S_{i-1}^B, c_i, F_{i-1}, \widetilde{R}). \tag{22}$$

For model training, according to the Eq. (20), the NMT model with bidirectional target attention is trained on a set of training examples $\{(X^n, Y^n)\}_{n=1}^N$:

$$\begin{aligned} L(\theta_3) = -\frac{1}{N}\Big\{ &\sum_{n=1}^{N}\sum_{i=1}^{I_y} \log P(y_i^n|y_{<i}^n, X^n; \theta_3) \\ &+ \sum_{n=1}^{N}\sum_{i=1}^{I_y} \log \overleftarrow{P}(\overleftarrow{y_i^n}|\overleftarrow{y_{>i}^n}, X^n; \theta_3)\Big\}. \end{aligned} \tag{23}$$

Finally, there is an available NMT model with bidirectional target attention parameterized by θ_3, as shown in Fig. 1 (c).

Transformer: the bidirectional model we propose can also be used in the Transformer to get more future target-side information. Since the structure of the Transformer only considers the forward target-side information, the influence of the future target-side information on the translation is not considered. Therefore, we add a reverse decoder module to the original transformer structure, as shown in Fig. 2, which

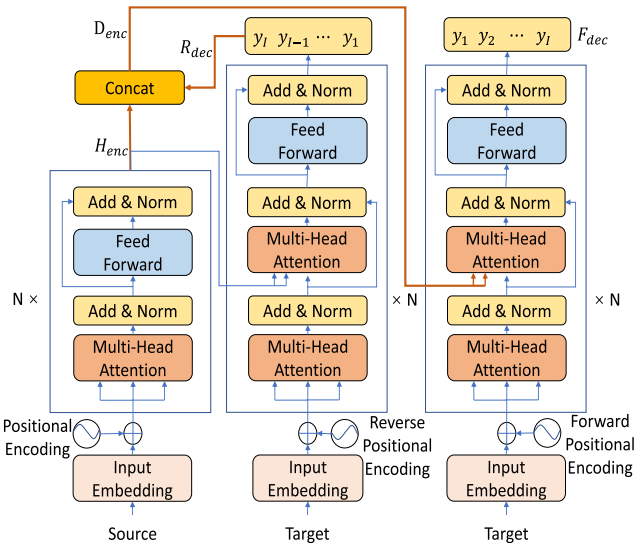

Fig. 2 Transformer with bidirectional target-attention model.

simultaneously applies both historical and future information when generating translations. In details, H_{enc}, R_{dec} and F_{dec} are the representations of the encoder, the reverse decoder and the forward decoder. R_{dec} can be computed as follows, similar to Eq. (8):

$$\begin{aligned} \overleftarrow{H}_{dec} &= \text{Attention}(\overleftarrow{Q}_y, \overleftarrow{K}_y, \overleftarrow{V}_y), \\ R_{dec} &= \text{Attention}(\overleftarrow{H}_{dec}, H_{enc}, H_{enc}), \end{aligned} \tag{24}$$

where $\overleftarrow{Q}_y = \overleftarrow{K}_y = \overleftarrow{V}_y$ are a reverse target sequence $\overleftarrow{Y}$. Mikolov *et al.* [9] use concatenation as the method to combine the sentence vectors to strengthen the capacity of representation. We also use the same method to combine H_{enc} and R_{dec}:

$$D_{enc} = \text{Concat}(H_{enc}, R_{dec}). \tag{25}$$

Finally, D_{enc} is added into forward context attention layer to get translation. In this way, Transformer can have the ability to use future target-side information. Which can be computed as follow:

$$F_{dec} = \text{Attention}(H_{dec}, D_{enc}, D_{enc}). \tag{26}$$

Based on the Eq. (9), our final loss is also composed of two parts, the formula is as follows:

$$\begin{aligned} L(\theta_4) = -\frac{1}{N}\Big\{ &\sum_{n=1}^{N}\sum_{i=1}^{I_y} \log P(y_i^n | y_{<i}^n, D_{enc}, F_{dec}; \theta_4) \\ &+ \sum_{n=1}^{N}\sum_{i=1}^{I_y} \log \overleftarrow{P}(\overleftarrow{y}_i^n | \overleftarrow{y}_{>i}^n, H_{enc}, R_{dec}; \theta_4)\Big\}. \end{aligned} \tag{27}$$

In the all two-pass decoding process, we have three steps. First, we use the reverse target attention layer with greedy search to sequentially generate reverse hidden states until the target-side start symbol <s> occurs with the highest probability. Then, we use all reverse hidden states to get the reverse target context R (with *average* operation in RNN). Finally, we add R into the forward decoder to find the best translation with *GRU* or *Attention* operation.

4. Experimentation

4.1 Experimental Settings

For Chinese-English translation, our training data for the translation task consists of 1.25M Chinese-English sentence pairs extracted from LDC corpora. The NIST02 test set is chosen as a development set, and the NIST03, NIST04, NIST05, NIST06 datasets are test sets. We use the case-insensitive *4*-gram NIST BLEU score as our evaluation metric [21]. The training data of English-German translation is from WMT 2015, which consists of 4.5M sentence pairs. We use byte-pair encoding [22] to segment words. The news-test-2016 was used as development set, the news-test-2014 and the news-test-2015 as test sets that are evaluated by SacreBLEU [23].

All NMT models are implemented in OpenNMT, including the proposed forward target attention based on RNN (**FTAtt-R**), reverse target attention based on RNN (**RTAtt-R**), bidirectional target attention based on RNN (**BiTAtt-R**) and bidirectioanl target attention based on Transformer (**BiTAtt-T**). On the Chinese-English and English-German translation, we limit the source and target vocabularies to the most frequent 32K words, and the maximum sentence length on both source and target sides to 50. In our three target attention models based on RNN, the dimensions of word embedding are 620, the size of the hidden layer is 1000 and the minibatch size is set as 80, the number of layers at the source and target of the RNN is 1, all the other settings are the same as in Bahdanau *et al.* [2]. We proposed **BiTAtt-T**, which consists of an encoder, a reverse decoder, and a forward decoder. Each of these three modules has 6 stacked layers of 512 neurons and the filter size of the layer is 2048. We set 512 neurons for the word embedding and minibatch size is also 512. About 200K minibathes are trained. All the other settings are the same as in Vaswani *et al.* [4]. We use an adam algorithm to train each model. We also re-implemented the following systems as our baselines:

PBSMT [19]: this is an open source hierarchical phrase-based SMT system with default configuration and a 4-gram language model.

ANMT [2]: this is an attention-based NMT with slight changes from OpenNMT.

ANMT(R2L): this is a variant of ANMT system with a right-to-left direction in target side.

ABDNMT [20]: this is an open source asynchronous bidirectional decoding for NMT system with default configuration.

TFMR: we implement the base Transformer model with a self-attention NMT [4].

Table 2 Translation results (BLEU score) for Chinese-English and English-German translation task. There are six existing experimental results to be shown. ReCons [11] is a encoder-decoder-reconstructor framework for NMT. MemDec [12] improves translation quality with external memory. NMT_{IA} [13] adds the last output information in the update of the attention weight. M-NMT [14] presents a memory-augmented NMT architecture, which stores knowledge about how words should be translated in a memory. DMAtt [15] incorporates word reordering knowledge into attention-based NMT. SDAtt [16] extend the local attention with syntax-distance constraint. BPEChar [17] is a character-level decoder without explicit segmentation for NMT. RecAtt [18] explicitly takes the attention history into consideration when generating the attention map. Avg means the average BLEU score on all test sets. "†": we proposed three target-attention methods based on RNN significantly better than ABDNMT, and our target-attention method based on Transformer significantly outperforms TFMR at significance level 0.05.

Type	Model	NIST					WMT		
		03	04	05	06	Avg	14	15	Avg
Report	ReCons [11]	N/A	N/A	34.88	35.19	N/A	N/A	N/A	N/A
	MemDec [12]	36.16	39.81	35.91	35.98	36.95	N/A	N/A	N/A
	NMT_{IA} [13]	35.09	37.73	35.53	34.32	35.67	N/A	N/A	N/A
	M-NMT [14]	34.00	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	DMAtt [15]	38.33	40.11	36.71	35.29	37.61	N/A	N/A	N/A
	SDAtt [16]	36.67	38.66	35.75	34.03	36.28	20.75	22.05	21.40
	BPEChar [17]	N/A	N/A	N/A	N/A	N/A	21.56	23.91	22.74
	RecAtt [18]	N/A	N/A	29.30	N/A	N/A	22.10	25.00	23.55
re-implement	PBSMT [19]	33.32	34.98	31.63	31.56	32.87	19.68	20.42	20.05
	ANMT [2]	36.42	39.33	35.37	35.56	36.67	22.42	25.13	23.76
	ANMT(R2L)	36.38	39.30	35.43	35.02	36.53	22.68	25.36	24.02
	ABDNMT [20]	39.84	42.16	38.67	38.19	39.72	23.46	26.13	24.80
	TFMR [4]	45.57	46.40	46.11	44.92	45.75	27.43	29.54	28.49
our RNN	**FTAtt-R**	40.35	42.58	39.62†	38.83†	40.35	23.62	26.35	24.99
	RTAtt-R	40.52†	42.72†	39.83†	38.97‡	40.51	23.80†	26.64†	25.22
	BiTAtt-R	**40.82**†	**43.09**†	**41.17**†	**39.35**†	**41.11**	**24.12**†	**26.81**†	**25.47**
our Transformer	**BiTatt-T**	**46.31**†	**47.15**†	**46.97**†	**45.71**†	**46.54**	**28.15**†	**30.13**†	**29.14**

4.2 Performance

Table 2 shows the performances measured in terms of BLEU score. **ABDNMT** outperforms the existing strong baseline **DMAtt** [15] by 2.1 BLEU points. **ANMT**, **ANMT(R2L)**, and **ABDNMT** outperform **PBNMT** by 3.8, 3.7, and 6.9 BLEU points respectively, indicating that **ANMT**, **ANMT(R2L)** and **ABDNMT** are stronger baselines.

With respect to BLEU scores, both of **RTAtt-R** and **FTAtt-R** have improved translation accuracy by 0.6 and 0.8 BLEU points on average over **ABDNMT**. Particularly, **BiTAtt-R** gets the most remarkable promotion, which beats the baseline **ABDNMT** with averaged 1.4 BLEU score on all test sets. This means that both forward and reverse target-attention information can work together well. Besides, our bidirectional target-attention model was successfully applied in the Transformer and achieved significant improvement of 0.8 BLEU points.

The proposed method gains similar improvements on English-German translation task. In addition, the performances of the proposed methods outperform the results in the existing works in both tasks.

5. Analysis

As the proposed three models achieve significant improvement over baseline, we further look at our models to explore how the target-side relationship plays a role in translation.

5.1 Efficiency Analysis

In Table 2, we analyze the efficiency of the proposed method. In RNN based NMT, compared to the ANMT, BiTAtt-R increases approximately 49% parameters and decrease approximately 57% training and 14% decoding speed. However, compared with ABDNMT, BiTAtt-R uses fewer parameters and is much faster in training and decoding.

In transformer based NMT, compared with TFMR(base), BiTAtt-T increases approximately 44% parameters and decreases approximately 36% training and 9% decoding speed. Compared with TFMR(big), BiTAtt-T just contains 40% the parameters. However, BiTAtt-T achieves similar performance with TFMR(big) and is much faster than TFMR(big).

The above empirical finds indicate that the improvement of the proposed methods does come not from more parameters. In the all two-pass decoding process, specifically, the decoding time of our system does not increase significantly. This is mainly because we use greedy search to generate reverse target hidden states in the first pass reverse decoding process, and employ beam search method (beam-size=10) the same as standard ANMT and Transformer in the second forward decoding process. This method is more time consuming than the greedy method, which is about 10 times.

Table 3 The efficiency analysis on English-German translation task. TFMR(big) differs TFMR(base) at the layer size (1024 vs 512) and the attention head number (16 vs 8). We have a single GPU device P100 to train/decode these models. The beamsize is set to 10 for decoding.

Type	Model	BLEU WMT14	Params	Speed (tokens/s)	
				Train	Decode
RNN	ANMT	22.42	84.4M	8200	214
	ABDNMT	23.46	130.0M	2300	97
	BiTAtt-R	24.12	125.4M	3500	183
Transformer	TFMR(base)	27.43	78.3M	10200	154
	TFMR(big)	28.26	282.8M	4500	99
	BiTAtt-T	28.15	113.0M	6500	140

Table 4 Chinese-English translation results of bidirectional target-attention model.

Source	新来 的 25 人 将 在 首都 汉城 附近 的 一 个 政府 收容所 上 适应 课 。
Reference	the 25 new arrivals will take adjustment lessons **in a government shelter** near the capital **city of seoul** .
TFMR(Base)	newly arrived 25 people will adapt to the course in a government **office** near the capital.
BiTAtt-T	newly arrived 25 people will adapt to course **in a government shelter** near the capital **city of seoul** .
Source	枪手 被 警方 击毙 。
Reference	the gunman was **shot** to **death** by **the police** .
TFMR(Base)	the gunmen were **shot to the police** .
BiTAtt-T	the gunmen were **shot** to **death** by **the police** .

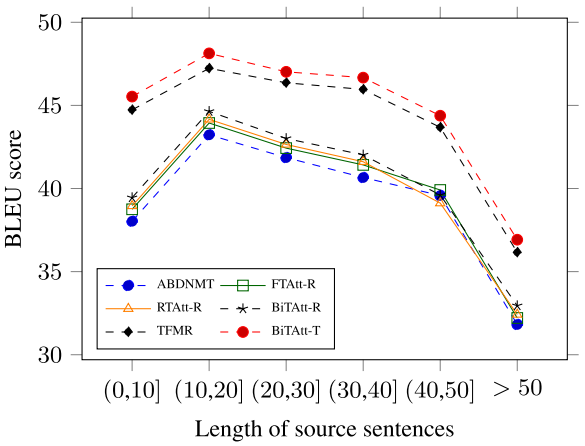

Fig. 3 BLEU score of generated translations with respect to the lengths of the input sentences on Chinese-English translation task.

5.2 Effects on Long Sentences

Following Bahdanau *et al.* [2], we group sentences of similar lengths together and compute BLEU score and averaged length of translation for each group, as shown in Fig. 3. It shows that the proposed FTAtt-R, RTAtt-R, BiTAtt-R, and BiTAtt-T outperform the baseline ABDNMT and TFMR over sentences with all different lengths respectively. We think the proposed target-attention can more effectively capture relationship among target words to improve target word prediction than the existing single decoder hidden state, which is in line with the effectiveness of target-side relationship found by Wu *et al.* [24].

Cho *et al.* [25] and Tu *et al.* [26] show that the performance of Groundhog drops rapidly when the length of the input sentence increases. Our results confirm these findings. It also shows that the performance drops substantially when the length of the input sentences increases, and thus faces a serious under-translation problem. It can be seen from the right side of Fig. 3, NMT systems tend to perform worse for long input sentences. We think the problem is that the maximum length limit of the source sentence is set to 50. For over 50 lengths of source sentences, our proposed NMT systems also have the lower performance, but still, exceed baselines in all groups. Our models relieve the under-translation problem to a certain extent.

5.3 Analysis on Translation Quality

Table 3 shows the translation examples. In the TFMR(Base), "收容所" is incorrectly translated into "office", instead of "shelter". According to the parse tree in the reference generated by the Stanford parser, the "shelter" has a forward relationship on the "in a government" and a reverse relationship on the "city". The "in a government" and "city" are very informative for correctly translate "收容所" to "shelter", but both of them are far away from "shelter" such that it is not easy to be captured by the TFMR(Base). Besides, BiTAtt-T correctly translated "汉城 (首尔 now)" into "Seoul", while TFMR(Base) ignores it. This information is considered in our bidirectional model which can solve the problem of error and under translation to a certain extent.

5.4 Analysis on Target-Side Alignment

Figure 4 shows the attention alignments for the translation example in Table 3. The BiTAtt-T does meet the expectation: the self-alignment in the target can capture the target-side relationship among the target words. We can find some phenomena to prove our method is valid. The words "death" and the forward word "to", "shot" have strong relevance. At the same time, the word "by" and the reverse word "police" have some correlation. While generating the current word,

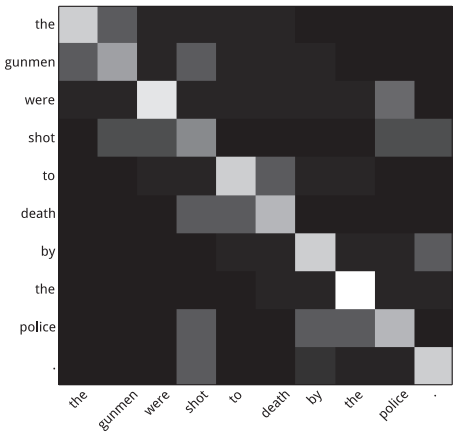

Fig. 4 Target alignment of BiTAtt-T Model.

the forward and the reverse information play a syntactic role. This example demonstrates that the proposed method can learn target-side relationship to help translate.

6. Related Work

In this section, we briefly review previous studies that are related to our work. Here we divide previous work into three categories: language model, attention mechanism, and target direction.

6.1 Language Model

In conventional SMT, the language model plays an important role. The application of neural networks to machine translation was restricted to extending standard machine translation tools for rescoring translation hypotheses or reranking n-best lists [27]–[31]. However, in the NMT system, the language model is usually replaced implicitly with an RNN model. Gülçehre *et al.* [32] proposed a method which integrates a language model into an attention-based NMT system. They can make full use of semantic information on the target-side. Similar to the language model, our methods force on the relationship between target words including forward and reverse. By using it effectively, we can improve the quality of the translation.

6.2 Attention Mechanism

Recent advance towards of NMT has achieved great success [2], [33]. In the NMT system, attention mechanism is a very effective and important method which learns to align and translate at the same time. It has greatly improved the performance of translation. On this basis, there are many interesting and effective methods [5], [16], [26], [34], [35] which have been proposed in improving attention mechanism of the NMT system. Luong *et al.* [5] proposed global attention model and local attention model, further compare several different scoring functions of the attention weight. Tu *et al.* [26] presented a coverage vector to keep track of the attention history and promote the attention mechanism to focus on more untranslated source words. Chen *et al.* [16], [34] proposed a double context method by two attention mechanism to capture more source context information for translation prediction. Our work has the same source attention mechanism, compared with the above models, the forward and the reverse target attention are also imported, which can help to produce a more smooth translation. Recently, the stacked self-attention layers were introduced in the Transformer model [4] and has significantly improved state-of-the-art in NMT. The difference was that we proposed reverse decoding and bidirectional decoding focus on the sentence-level instead of the monodirectional decoding in the Transformer. Specifically, our method simply adds reverse target-attention into the forward decoder to improve translation prediction which can be transferred to the other machine translation systems easily.

6.3 Target Direction

Target-directional neural network models have also been successfully employed in Devlin *et al.* [28]. However, their approach was concerned with feedforward networks. Sennrich *et al.* [36] attempted to re-rank the "left-to-right" decoding results by "right-to-left" decoding, resulting in diversified translation results. Similar in spirit to this, Li *et al.* [37] introduced a beam search algorithm which can be diversified by integrating bidirectional scores in re-ranking, or by adjusting the beam diversity with reinforcement learning [38]. Cheng *et al.* [39] proposed a bidirectional attention model for joint training, so as to keep consistent in two directions. Liu *et al.* [40] tried to jointly train by using two directional models and then search for target sequences which have support from both of the models in testing. Zhou *et al.* [41] also proposed a synchronous bidirectional decoding to produce better translation. It is notable that Xia *et al.* [42] and Zhang *et al.* [20] presented target attention models which are similar to us. However, the former does not consider reverse target semantic information, and the latter differs from ours in three aspects: (1) our models consider the forward target attention information and (2) our models generate the final translation with applying the forward and the reverse target attention information simultaneously. (3) Our models also give the Transformer the ability to get the future target attention information. Different from the previous studies, our proposed model takes full account of the forward and reverse target information, and combines them efficiently to help to generate the target sequence.

7. Conclusion

In this paper, we have presented three novel approaches that incorporate the whole relationship among target words into traditional NMT and Transformer with target-side attention models. The difference between the three models is the direction of the target-side relationship. Our bidirectional target-attention model can effectively learn both forward and

reverse target semantic information to help translation. Experiment results on Chinese-English and English-German translation have demonstrated the efficacy of the proposed models. We have also analyzed the translation behavior of our improved system against the state-of-the-art NMT baseline system from several perspectives, indicating that there is much room for NMT translation to be enhanced by more semantic information. Since the proposed models are a simple universal sequence-to-sequence framework, we can easily apply them to other sequence-to-sequence models and tasks in the future.

References

[1] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, pp.1700–1709, Association for Computational Linguistics, Oct. 2013.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 3rd International Conference on Learning Representations, San Diego, CA, USA, Conference Track Proceedings, May 2015.

[3] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y.N. Dauphin, "Convolutional sequence to sequence learning," Proceedings of the 34th International Conference on Machine Learning, ed. D. Precup and Y.W. Teh, International Convention Centre, Sydney, Australia, pp.1243–1252, Aug. 2017.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, and I. Polosukhin, "Attention is all you need," Proceedings of Advances in Neural Information Processing Systems 30, ed. I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, pp.5998–6008, Curran Associates, Inc., Dec. 2017.

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

[6] D. Xiong, M. Zhang, and H. Li, "Enhancing language models in statistical machine translation with backward n-grams and mutual information triggers," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp.1288–1297, Association for Computational Linguistics, June 2011.

[7] J. Mitchell and M. Lapata, "Composition in distributional models of semantics," Cognitive Science, vol.34, no.8, pp.1388–1429, 2010.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," CoRR, vol.abs/1301.3781, 2013.

[9] Q.V. Le and T. Mikolov, "Distributed representations of sentences and documents.," Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, JMLR Workshop and Conference Proceedings, vol.32, pp.1188–1196, June 2014.

[10] R. Wang, A. Finch, M. Utiyama, and E. Sumita, "Sentence embedding for neural machine translation domain adaptation," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada, pp.560–566, July 2017.

[11] Z. Tu, Y. Liu, L. Shang, X. Liu, and H. Li, "Neural machine translation with reconstruction," Proceedings of the 31st AAAI Conference on Artificial Intelligence, Feb. 4–9, 2017, San Francisco, California, USA., pp.3097–3103, Feb. 2017.

[12] M. Wang, Z. Lu, H. Li, and Q. Liu, "Memory-enhanced decoder for neural machine translation," Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp.278–286, Association for Computational Linguistics, 2016.

[13] F. Meng, Z. Lu, H. Li, and Q. Liu, "Interactive attention for neural machine translation," Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, pp.2174–2185, The COLING 2016 Organizing Committee, Dec. 2016.

[14] Y. Feng, S. Zhang, A. Zhang, D. Wang, and A. Abel, "Memory-augmented neural machine translation," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, pp.1390–1399, Association for Computational Linguistics, Sept. 2017.

[15] J. Zhang, M. Wang, Q. Liu, and J. Zhou, "Incorporating word reordering knowledge into attention-based neural machine translation," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, pp.1524–1534, Association for Computational Linguistics, July 2017.

[16] K. Chen, R. Wang, M. Utiyama, E. Sumita, and T. Zhao, "Syntax-directed attention for neural machine translation," Proc. 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, pp.4792–4799, Feb. 2018.

[17] J. Chung, K. Cho, and Y. Bengio, "A character-level decoder without explicit segmentation for neural machine translation," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, pp.1693–1703, Association for Computational Linguistics, Aug. 2016.

[18] Z. Yang, Z. Hu, Y. Deng, C. Dyer, and A. Smola, "Neural machine translation with recurrent attention modeling," Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, pp.383–387, Association for Computational Linguistics, April 2017.

[19] C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman, and P. Resnik, "cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models," Proceedings of the Association for Computational Linguistics 2010 System Demonstrations, Uppsala, Sweden, pp.7–12, Association for Computational Linguistics, 2010.

[20] X. Zhang, J. Su, Y. Qin, Y. Liu, R. Ji, and H. Wang, "Asynchronous bidirectional decoding for neural machine translation," Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, pp.5698–5705, AAAI Press, Feb. 2018.

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

[22] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, pp.1715–1725, Association for Computational Linguistics, Aug. 2016.

[23] M. Post, "A call for clarity in reporting BLEU scores," Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels, pp.186–191, Association for Computational Linguistics, Oct. 2018.

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

[25] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation,"

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, pp.1724–1734, Association for Computational Linguistics, Oct. 2014.
[26] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Modeling coverage for neural machine translation," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, pp.76–85, Association for Computational Linguistics, Aug. 2016.
[27] H. Schwenk, "Continuous space translation models for phrase-based statistical machine translation," Proceedings of the 24th International Conference on Computational Linguistics: Posters, Mumbai, India, pp.1071–1080, The COLING 2012 Organizing Committee, Dec. 2012.
[28] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, "Fast and robust neural network joint models for statistical machine translation," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, pp.1370–1380, Association for Computational Linguistics, June 2014.
[29] R. Wang, H. Zhao, B.-L. Lu, M. Utiyama, and E. Sumita, "Bilingual continuous-space language model growing for statistical machine translation," Trans. Audio, Speech and Lang. Proc., vol.23, no.7, pp.1209–1220, July 2015.
[30] R. Wang, M. Utiyama, I. Goto, E. Sumita, H. Zhao, and B.L. Lu, "Converting continuous-space language models into n-gram language models with efficient bilingual pruning for statistical machine translation," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol.15, no.3, pp.11:1–11:26, Jan. 2016.
[31] K. Chen, T. Zhao, M. Yang, L. Liu, A. Tamura, R. Wang, M. Utiyama, and E. Sumita, "A neural approach to source dependence based context model for statistical machine translation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.26, no.2, pp.266–280, Feb. 2018.
[32] Ç. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," CoRR, vol.abs/1503.03535, March 2015.
[33] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," Proceedings of Transactions of the Association for Computational Linguistics, vol.4, pp.371–383, Aug. 2016.
[34] K. Chen, R. Wang, M. Utiyama, L. Liu, A. Tamura, E. Sumita, and T. Zhao, "Neural machine translation with source dependency representation," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, pp.2846–2852, Association for Computational Linguistics, Sept. 2017.
[35] S. Kuang, J. Li, A. Branco, W. Luo, and D. Xiong, "Attention focusing for neural machine translation by bridging source and target embeddings," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.1767–1776, Association for Computational Linguistics, July 2018.
[36] R. Sennrich, B. Haddow, and A. Birch, "Edinburgh neural machine translation systems for wmt 16," Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, Berlin, Germany, pp.371–376, Association for Computational Linguistics, Aug. 2016.
[37] J. Li and D. Jurafsky, "Mutual information and diverse decoding improve neural machine translation," CoRR, vol.abs/1601.00372, March 2016.
[38] J. Li, W. Monroe, and D. Jurafsky, "A simple, fast diverse decoding algorithm for neural generation," CoRR, vol.abs/1611.08562, Dec. 2016.
[39] Y. Cheng, S. Shen, Z. He, W. He, H. Wu, M. Sun, and Y. Liu, "Agreement-based joint training for bidirectional attention-based neural machine translation," Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, USA, pp.2761–2767, AAAI Press, July 2016.
[40] L. Liu, A. Finch, M. Utiyama, and E. Sumita, "Agreement on target-bidirectional lstms for sequence-to-sequence learning," Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, Arizona, pp.2630–2637, AAAI Press, July 2016.
[41] L. Zhou, J. Zhang, and C. Zong, "Synchronous bidirectional neural machine translation," Transactions of the Association for Computational Linguistics, vol.7, pp.91–105, 2019.
[42] Y. Xia, F. Tian, T. Qin, N. Yu, and T.-Y. Liu, "Sequence generation with target attention," Proceedings of the 2017 Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Skopje, Macedonia, pp.816–831, Springer, Cham, Sept. 2017.

Mingming Yang received the B.S. degree in computer science from Harbin University of Science and Technology, China in 2010 and the M.S. degree in computer science from Harbin University of Science and Technology, China in 2013. He is currently a Ph.D candidate in Harbin Institute of Technology from 2013. His research interests include machine translation and natural language processing.

Min Zhang is a distinguished professor in the School of Computer Science and Technology, Soochow University, Suzhou. He received his Bachelor degree and Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, in 1991 and 1997, respectively. His current research interests include machine translation, natural language processing, and artificial intelligence.

Kehai Chen received the B.S. degree in computer science from Xi'an University of Technology in 2010, the M.S. degree in computer science from University of Chinese Academy of Sciences in 2013, and the Ph.D. degree in computer science from Harbin Institute of Technology in 2018. He was an internship researcher fellow in National Institute of Information and Communications Technology, Japan since 2017. He is a researcher in National Institute of Information and Communications Technology, Japan since 2018. His research interests include machine translation and natural language processing.

Rui Wang received the B.S. degree from Harbin Institute of Technology in 2009, the M.S. degree from Chinese Academy of Sciences in 2012 and the Ph.D. degree in Shanghai Jiao Tong University in 2016, all of which are in computer science. He was a Joint Ph.D. in Centre Nationnal de la Recherche Scientifique, France in 2014. He is a researcher in National Institute of Information and Communications Technology, Japan since 2016. His research interests include machine translation and natural language processing.

Tiejun Zhao is a professor of School of Computer Science and Technology, Harbin Institute of Technology. His research interests include: natural language understanding, content-based web information processing, and applied artificial intelligence. He has published 3 academic books and 60 papers on journals and conferences in recent 3 years. Prof. Zhao has been a PC member on ACL, COLING in current 5 years and was also assigned an MT Track Co-chair on COLING 2014.