Artificial Intelligence 277 (2019) 103168

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Exploiting reverse target-side contexts for neural machine translation via asynchronous bidirectional decoding

Jinsong Su [a], Xiangwen Zhang [a], Qian Lin [a], Yue Qin [a], Junfeng Yao [a], Yang Liu [b,*]

[a] Xiamen University, Xiamen 361005, China
[b] Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 8 August 2018
Received in revised form 28 March 2019
Accepted 27 August 2019
Available online 4 September 2019

Keywords:
Neural machine translation
Bidirectional decoding
Attention mechanism

ABSTRACT

Based on a unified encoder-decoder framework with attentional mechanism, neural machine translation (NMT) models have attracted much attention and become the mainstream in the community of machine translation. Generally, the NMT decoders produce translation in a left-to-right way. As a result, only left-to-right target-side contexts from the generated translations are exploited, while the right-to-left target-side contexts are completely unexploited for translation. In this paper, we extend the conventional attentional encoder-decoder NMT framework by introducing a backward decoder, in order to explore asynchronous bidirectional decoding for NMT. In the first step after encoding, our backward decoder learns to generate the target-side hidden states in a right-to-left manner. Next, in each timestep of translation prediction, our forward decoder concurrently considers both the source-side and the reverse target-side hidden states via two attention models. Compared with previous models, the innovation in this architecture enables our model to fully exploit contexts from both source side and target side, which improve translation quality altogether. We conducted experiments on NIST Chinese-English, WMT English-German and Finnish-English translation tasks to investigate the effectiveness of our model. Experimental results show that (1) our improved RNN-based NMT model achieves significant improvements over the conventional RNNSearch by 1.44/-3.02, 1.11/-1.01, and 1.23/-1.27 average BLEU and TER points, respectively; and (2) our enhanced Transformer outperforms the standard Transformer by 1.56/-1.49, 1.76/-2.49, and 1.29/-1.33 average BLEU and TER points, respectively. We released our code at https://github.com/DeepLearnXMU/ABD-NMT.

1. Introduction

With the rapid advancement of deep learning in various NLP tasks, the studies of machine translation have shifted from statistical machine translation (SMT) to neural machine translation (NMT) [10,26,5]. Unlike the conventional SMT [13,4] that needs to explicitly design features to learn translation knowledge, NMT aims at constructing an end-to-end encoder-decoder framework based on neural networks to model the entire translation process. Among these models, the

* Corresponding author.
E-mail addresses: jssu@xmu.edu.cn (J. Su), xwzhang@stu.xmu.edu.cn (X. Zhang), linqian17@stu.xmu.edu.cn (Q. Lin), qinyue@stu.xmu.edu.cn (Y. Qin), yao0010@xmu.edu.cn (J. Yao), liuyang2011@tsinghua.edu.cn (Y. Liu).

https://doi.org/10.1016/j.artint.2019.103168
0004-3702/

recent developed attentional NMT [2] has quickly become the most commonly-used model, due to its excellent capability in capturing long-distance dependencies for translation.

In general, most NMT decoders are based on recurrent neural network (RNN) or autoregressive Transformer network, where translation procedures are conducted in a left-to-right manner. Although these decoders are able to capture and memorize unbounded target words generated previously, the reverse target-side contexts are completely unexploited for translation. As a result, if errors occurred in previous translation predictions, the quality of subsequent translations is very likely to be undermined due to the noisy left-to-right encoded target-side contexts. Intuitively, the reverse target-side contexts also play an important role in translation predictions, for they not only provide complementary signals but also induce different biases to NMT model [7]. For example, let us consider the following two candidate translations respectively produced by the NMT system with left-to-right and right-to-left decoding:

- **Source**: rì fángwèitīng zhǎngguān : bú wàng jūnguó lìshǐ zūnzhòng línguó zūnyán
- **Reference**: *japan defense chief : never forget militaristic history , respect neighboring nations ' dignity*
- **NMT(left-to-right decoding)**: *japan 's defense agency chief : death of militarism respects its neighbors ' dignity*
- **NMT(right-to-left decoding)**: *japanese defense agency has never forgotten militarism 's history to respect the dignity of neighboring countries*

In this example, the latter half of the Chinese sentence is accurately translated by the NMT system equipped with right-to-left decoder, while the conventional NMT system generates misinterpreted translation. Thus, how to effectively encode reverse target-side contexts and use them as supplement to left-to-right target-side contexts is crucial to improve the translation performance of NMT.

To achieve this goal, many researchers focused on introducing bidirectional decoding in NMT [15,19,7]. For the sake of selecting a translation with both appropriate prefixes and suffixes, most of them used bidirectional decoding scores together to re-rank candidate translations. However, such methods are always accompanied with some drawbacks that limit the ability of NMT to model target-side contexts. On the one hand, the limited search space and search errors of beam search lead to unsatisfactorily generated 1-best translation, which fails to further provide sufficient complementary contexts for the decoder in the other direction. On the other hand, due to the independence between two unidirectional decoders in translation process, the unidirectional decoder lacks of capability to fully exploit target-side contexts generated by the other decoder, leading to the undesirably generated candidate translations. Therefore, how to effectively excavate the effect of bidirectional decoding on NMT still deserves further study.

In order to fully exploit reverse target-side contexts for NMT, in this paper, we propose a novel **NMT** framework with **A**synchronous **B**idirectional **D**ecoding (ABD-NMT), which significantly extends the conventional attentional encoder-decoder NMT framework by introducing a backward decoder. The framework representation of our model is illustrated in Fig. 1. Along with our novel asynchronous bidirectional decoders, the proposed model remains an end-to-end attentional NMT framework, which mainly includes three components: 1) an encoder that embeds the input source sentence into bidirectional hidden states; 2) a backward decoder which generates translation in the right-to-left manner and the resulting hidden states encode the reverse target-side contexts; 3) a forward decoder that generates the final translation from left to right, where two attention models are simultaneously employed to consider both the source-side bidirectional and target-side reverse hidden state vectors for translation predictions. Compared with the previous related NMT models, our model has significant advantages as follows: 1) The hidden state vectors produced by the backward decoder are able to encode semantics of potential hypotheses, serving as extra target-side contexts utilized by the following forward decoder during translation procedure. 2) By integrating right-to-left target-side context modeling and left-to-right translation generation into an end-to-end joint framework, our model alleviates the error propagation of reverse target-side context modeling to some extent.

The major contributions of this paper are concluded as follows:

- Through the thorough analyses, we point out the existing drawbacks of studies on the conventional NMT and the NMT with bidirectional decoding.
- We introduce asynchronous bidirectional decoding into NMT, where a backward decoder is used to encode the left-to-right target-side contexts, as a supplement to the conventional context modeling mechanism. To the best of our knowledge, our work is the first attempt to investigate the effectiveness of the end-to-end attentional NMT model with asynchronous bidirectional decoders.
- Experimental results on NIST Chinese-English, WMT English-German and Finnish-English translation tasks show that our model achieves significant improvements over the conventional NMT model.

This work has been presented in our previous conference paper [31]. In this article, we make the following significant extensions to our previous work:

- Through various groups of experiments, we deeply analyze how our proposed model outperforms other baselines. Finally, we believe our model is able to exploit both the reverse target-side context and two-pass decoding to benefit NMT models.

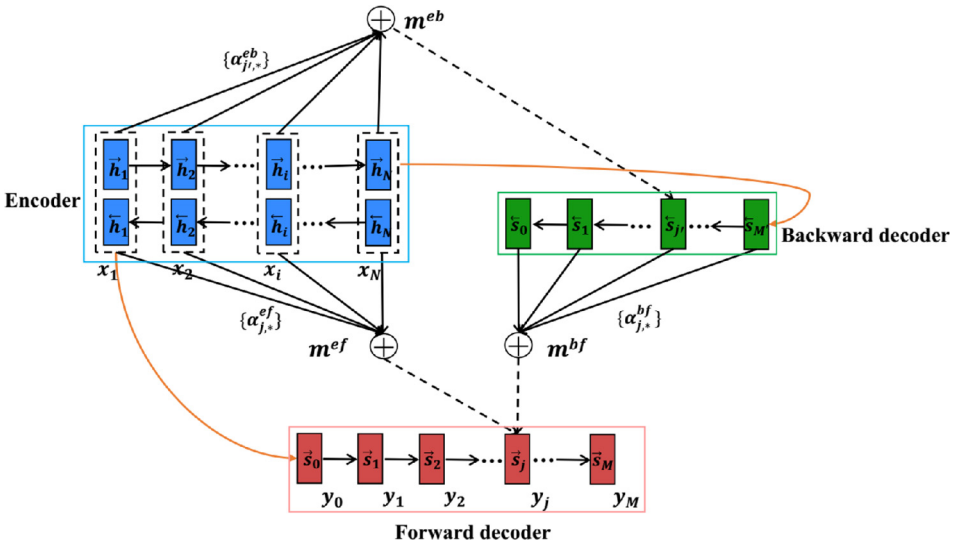

Fig. 1. The architecture of the proposed NMT model. Note that the forward decoder directly attends to the reverse hidden state sequence rather than the word sequence produced by the backward decoder.

- We apply our method into the state-of-the-art Transformer [27] and demonstrate its effectiveness on this NMT framework.
- We conduct more experiments on Finnish-English translation to further investigate the effectiveness of our model on more language pairs.
- We provide more details of our model, such as the numbers of model parameters, as well as model running time.

The remainder of this paper is organized as follows. Section 2 summarizes the related work and highlights the differences of our model from previous studies; Section 2 elaborates our proposed model, including details on the model framework, objective function and training procedure; Experimental results are reported and analyzed in Section 4; Finally, we conclude with future work in Section 5.

2. Related work

Bidirectional decoding has always been a research focus in machine translation. Here we summarize and briefly introduce the related work in SMT and NMT, respectively.

In SMT, many approaches related with backward language model or right-to-left decoding have been explored to capture reverse target-side contexts for translation. In this respect, Watanabe and Sumita [28] first explored two decoding methods: one is the right-to-left decoding based on the left-to-right beam search, which generates outputs from the end of a sentence; the other is the bidirectional decoding method which decodes in both the left-to-right and right-to-left directions and then merges the two hypothesized partial sentences into one. Finch and Sumita [6] investigated the effects on phrase-based SMT of three different decoding strategies: forward, reverse and bidirectional. Moreover, they integrated the backward language model into the reverse translation decoder. Aside from left-to-right decoding, Zhang et al. [29] studied the effects of multiple decomposition structures as well as dynamic bidirectional decomposition on SMT.

When it comes to NMT, the dominant RNN-based NMT models usually perform translation in a left-to-right manner, leading to the limitation that the reverse target-side contexts are underutilized. To alleviate this drawback, Liu et al. [15] first jointly train both directional LSTM models, and then in testing they try to search for target-side translations that are encouraged by both directional models. Likewise, Sennrich et al. [19] rescored the left-to-right decoding results by right-to-left decoding, which brings about diversified translation results. Recently, Hoang et al. [7] proposed an approximate inference framework based on continuous optimization that enables bidirectional decoding. Finally, please note that pre-translation [16,32] and neural automatic post-editing [17,9] for NMT are also related to our work, because our model also involves two passes of translation.

Overall, the most relevant models include [15,19,7,32,17,9]. Our model can be significantly distinguished from these studies in the following aspects: 1) We hold different motivations from these studies. Specifically, in this work, we aim to improve NMT with left-to-right decoding by fully exploiting the reverse target-side contexts, which are encoded within right-to-left hidden state vectors. In contrast, Liu et al. [15], Sennrich et al. [19], Hoang et al. [7] explored how to use bidirectional decoding scores to generate better translations, both Niehues et al. [16] and Zhou et al. [32] committed to combine the advantages of both NMT and SMT, and in the work of [17,9], multiple neural architectures were exploited for the task of automatic post editing of machine translation output. 2) Instead of using the raw best output of machine

translation systems as implemented in [16,32,17,9], our model exploits reverse target-side context by considering the hidden state vectors in right-to-left direction. 3) Our model is an end-to-end NMT model, while the bidirectional decoders adopted in [15,19,7] were independent from each other, and the component used to produce the raw translation was independent from the NMT model in [16,32,17,9]. In addition, Serban et al. [21] introduced a backward RNN to refine the forward RNN. But their approach was only applied in the training procedure, which is different from ours where both the forward and backward RNNs were simultaneously used for sequence generation.

3. Our proposed model

As mentioned above, our model mainly consists of three components: 1) a neural encoder with parameter set $\boldsymbol{\theta_e}$; 2) a neural backward decoder with parameter set $\boldsymbol{\theta_b}$; and 3) a neural forward decoder with parameter set $\boldsymbol{\theta_f}$, which will be elaborated in the following subsections.

It should be noted that in this work we employ Gated Recurrent Unit (GRU) [5] to construct the encoder and decoders, because it has relatively few required parameters and has been widely used in the community of NMT. Nevertheless, our model is also applicable to models related to the RNN with other units, such as Long Short-Term Memory (LSTM) [8].

3.1. The neural encoder

Identical to the dominant NMT model, our neural encoder is also modeled using a bidirectional RNN.

Firstly, a source sentence $\mathbf{x}$=$x_1, x_2 ... x_N$ is read by the forward RNN in a left-to-right order. In this process, we employ a recurrent state transition function $\phi(\cdot)$ to generate the semantic representation of the word sequence $\mathbf{x}_{1:i}$ as $\overrightarrow{h}_i$=$\phi(\overrightarrow{h}_{i-1}, x_i)$ at each timestep. Similarly, the backward RNN reversely scans the source sentence and produces the semantic representation $\overleftarrow{h}_i$ of the word sequence $\mathbf{x}_{i:N}$. Finally, the hidden states of these two RNNs are concatenated to form an annotation sequence $\mathbf{h} = \{h_1, h_2, ...h_i..., h_N\}$, where $h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i]$ encodes the semantic information about the i-th word considering all the other surrounding words in the source sentence.

In our model, these annotations will provide source-side contexts for not only the forward decoder but also the backward one via different attention models.

3.2. The neural backward decoder

The neural backward decoder of our model is also similar to that of dominant NMT model, with the difference that it performs right-to-left decoding.

Given the source-side annotations and all target words generated previously, the backward decoder models how to reversely produce the next target word. Applying this decoder, we calculate the conditional probability of the reverse translation $\overleftarrow{\mathbf{y}}$=$(y_0, y_1, y_2, ..., y_{M'})$ as follows

$$\begin{aligned} P(\overleftarrow{\mathbf{y}}|\mathbf{x};\boldsymbol{\theta_e},\boldsymbol{\theta_b}) &= \prod_{j'=0}^{M'} P(y_{j'}|\mathbf{y}_{>j'},\mathbf{x};\boldsymbol{\theta_e},\boldsymbol{\theta_b}) \\ &= \prod_{j'=0}^{M'} \overleftarrow{g}(y_{j'+1}, \overleftarrow{s}_{j'}, m^{eb}_{j'}), \end{aligned} \tag{1}$$

where $\overleftarrow{g}(\cdot)$ is a non-linear function, $\overleftarrow{s}_{j'}$ and $m^{eb}_{j'}$ denote the decoding state and the source-side context vector at the j'-th time step, respectively, and M' indicates the length of the reverse hidden state sequence.

As for $\overleftarrow{s}_{j'}$ and $m^{eb}_{j'}$, we use the GRU activation function $\overleftarrow{s}_{j'}$=$f(\overleftarrow{s}_{j'+1}, y_{j'+1}, m^{eb}_{j'})$ to compute $\overleftarrow{s}_{j'}$, and introduce an **encoder-backward decoder attention model** to define $m^{eb}_{j'}$ as the weighted sum of the source-side semantic annotations $\{h_i\}$:

$$m^{eb}_{j'} = \sum_{i=1}^{N} \alpha^{eb}_{j',i} \cdot h_i, \tag{2}$$

$$\alpha^{eb}_{j',i} = \frac{\exp(e^{eb}_{j',i})}{\sum_{i'=1}^{N} \exp(e^{eb}_{j',i'})}, \tag{3}$$

$$e^{eb}_{j',i} = (v^{eb}_a)^\top \tanh(W^{eb}_a \overleftarrow{s}_{j'+1} + U^{eb}_a h_i), \tag{4}$$

where v^{eb}_a, W^{eb}_a and U^{eb}_a are the parameters of the encoder-backward decoder attention model. Through such implementations, the related source words will be automatically captured to predict target words from the opposite direction.

By introducing this backward decoder, our NMT model is able to encode the reverse target-side contexts for the subsequent left-to-right translation predictions. More importantly, compared with the target word sequence generated by the backward decoder, the corresponding target-side hidden states $\overleftarrow{\mathbf{s}}$ provides richer reverse target-side contexts to the forward decoder for the further use.

3.3. The neural forward decoder

The neural forward decoder of our model is an extension of dominant NMT decoder. It performs decoding in a left-to-right manner under the semantic guides of source-side and reverse target-side contexts, which are separately captured by two different attention models, respectively.

The forward decoder is trained to sequentially predict the next target word, whose inputs are from three aspects: (1) the source-side annotations of the encoder; (2) the reverse target-side hidden state sequence generated by the backward encoder, and (3) all target words generated by itself previously. Formally, the conditional probability of the translation $\mathbf{y}=(y_0, y_1, ..., y_M)$ is defined as follows:

$$\begin{aligned} P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta_e},\boldsymbol{\theta_b},\boldsymbol{\theta_f}) &= \prod_{j=0}^{M} P(y_j|\mathbf{y}_{<j},\mathbf{x};\boldsymbol{\theta_e},\boldsymbol{\theta_b},\boldsymbol{\theta_f}) \\ &= \prod_{j=1}^{M} g(y_{j-1}, s_j, m_j^{ef}, m_j^{bf}), \end{aligned} \tag{5}$$

where $g(\cdot)$ is a non-linear function, s_j is the decoding state, m_j^{ef} and m_j^{bf} denote the source-side and reverse target-side context vectors at the j-th timestep, respectively.

As illustrated in Fig. 1, we initialize the first hidden state s_0 of the forward decoder with the last hidden state of the reverse encoder at the last layer, which is represented as $\overleftarrow{h}_1$. Moreover, we employ two attention models to respectively produce the source-side and reverse target-side context vectors:

One is the **encoder-forward decoder attention model** which considers the source-side annotations. To be specific, we produce m_j^{ef} from the hidden states $\{h_i\}$ of the encoder as follows:

$$m_j^{ef} = \sum_{i=1}^{N} \alpha_{j,i}^{ef} \cdot h_i, \tag{6}$$

$$\alpha_{j,i}^{ef} = \frac{\exp(e_{j,i}^{ef})}{\sum_{i'=1}^{N} \exp(e_{j,i'}^{ef})}, \tag{7}$$

$$e_{j,i}^{ef} = (v_a^{ef})^{\top} \tanh(W_a^{ef} s_{j-1} + U_a^{ef} h_i), \tag{8}$$

where v_a^{ef}, W_a^{ef}, and U_a^{ef} are the parameters of the encoder-forward decoder attention model.

The other is the **backward decoder-forward decoder attention model** that focuses on reverse target-side hidden states. Likewise, we define m_j^{bf} as a weighted sum of the hidden states $\{\overleftarrow{s}_{j'}\}$ of the backward decoder. Formally, m_j^{bf} is calculated as follows:

$$m_j^{bf} = \sum_{j'=0}^{M'} \alpha_{j,j'}^{bf} \cdot \overleftarrow{s}_{j'}, \tag{9}$$

$$\alpha_{j,j'}^{bf} = \frac{\exp(e_{j,j'}^{bf})}{\sum_{j''=0}^{M'} \exp(e_{j,j''}^{bf})}, \tag{10}$$

$$e_{j,j'}^{bf} = (v_a^{bf})^{\top} \tanh(W_a^{bf} s_{j-1} + U_a^{bf} \overleftarrow{s}_{j'}), \tag{11}$$

where v_a^{bf}, W_a^{bf}, and U_a^{bf} are the parameters of the backward decoder-forward decoder attention model. Note that we directly choose hidden state sequence rather than word sequence to model the target-side contexts, due to the superiority of the former in avoiding negative effect of translation prediction errors to some extent.

Next, we incorporate the generated m_j^{ef} and m_j^{bf} into the GRU hidden unit of the forward decoder. Thus, the hidden state $\overrightarrow{s}_j$ of the forward decoder is formally computed by

$$\overrightarrow{s}_j = (1 - z_j^d) \circ \overrightarrow{s}_{j-1} + z_j^d \circ \tilde{s}_j,$$
$$\tilde{s}_j = \tanh(W^d v(y_{j-1}) + U^d [r_j^d \circ \overrightarrow{s}_{j-1}] + C^{ef} m_j^{ef} + C^{bf} m_j^{bf}), \tag{12}$$

where W^d, U^d, C^{ef}, and C^{bf} are the weight matrices, z_j^d and r_j^d are update and reset gates of GRU, respectively, depending on y_{j-1}, $\overrightarrow{s}_{j-1}$, m_j^{ef} and m_j^{bf}.

Finally, we further define the probability of y_j as

$$p(y_j|\mathbf{y}_{<j}, \mathbf{x}; \boldsymbol{\theta_e}, \boldsymbol{\theta_b}, \boldsymbol{\theta_f}) \propto \exp(g(y_{j-1}, \overrightarrow{s}_j, m_j^{ef}, m_j^{bf})), \tag{13}$$

where y_{j-1}, $\overrightarrow{s}_j$, m_j^{ef} and m_j^{bf} are concatenated and fed through a single feed-forward layer.

3.4. Model training

The objective function of our proposed model over the training corpus D={$(\mathbf{x}, \mathbf{y})$} is defined as

$$J(D; \boldsymbol{\theta_e}, \boldsymbol{\theta_b}, \boldsymbol{\theta_f}) = \frac{1}{|D|} \operatorname*{arg\,max}_{\boldsymbol{\theta_e}, \boldsymbol{\theta_b}, \boldsymbol{\theta_f}} \sum_{(\mathbf{x},\mathbf{y}) \in D} \{\lambda \cdot \log P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta_e}, \boldsymbol{\theta_b}, \boldsymbol{\theta_f}) + (1 - \lambda) \cdot \log P(\overleftarrow{\mathbf{y}}|\mathbf{x}; \boldsymbol{\theta_e}, \boldsymbol{\theta_b})\}, \tag{14}$$

where the reverse reference $\overleftarrow{\mathbf{y}}$ is obtained by inverting $\mathbf{y}$, and the hyper-parameter λ is used to balance the preference between the two terms.

In this objective function, the first term $\log P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta_e}, \boldsymbol{\theta_b}, \boldsymbol{\theta_f})$ models the translation procedure illustrated in Fig. 1. To ensure the consistency between model training and testing, we perform beam search to generate reverse hidden states $\overleftarrow{\mathbf{s}}$ when optimizing $\log P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta_e}, \boldsymbol{\theta_b}, \boldsymbol{\theta_f})$. In particular, to guarantee that the $\overleftarrow{\mathbf{s}}$ produced by beam search is of high quality, we further introduce the second term $\log P(\overleftarrow{\mathbf{y}}|\mathbf{x}; \boldsymbol{\theta_e}, \boldsymbol{\theta_b})\}$ to maximize the conditional likelihood of $\overleftarrow{\mathbf{y}}$, and use the hyper-parameter λ to balance the preference between the two terms. Due to the high time complexity required by beam search, we directly apply greedy search to implement right-to-left decoding, which proves to be sufficiently effective in our experiments. When translating the unseen sentence $\mathbf{x}$ using our proposed model, we first use the backward decoder with greedy search to sequentially generate $\overleftarrow{\mathbf{s}}$ until the target-side start symbol $\langle s \rangle$ occurs with the highest probability. Then, we apply beam search on the forward decoder to search the final translation with the maximal $\log P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta_e}, \boldsymbol{\theta_b}, \boldsymbol{\theta_f})$.

4. Experiments

To testify the effectiveness of the proposed model, we conducted experiments on NIST Chinese-English, WMT English-German and Finnish-English translation tasks.

4.1. Setup

For Chinese-English translation, the training set was from LDC corpus,[1] which includes 2M bilingual sentences with 54.8M Chinese words and 60.8M English words. For the validation set, we used NIST 2006 (MT06) dataset as our validation set and NIST 2002 (MT02), NIST 2003 (MT03), 2004 (MT04) and 2005 (MT05) as test sets.

For English-German translation, we used WMT 2015 training data that contains 4.46M sentence pairs with 116.1M English words and 108.9M German words. The validation set and the test set were news-test 2013 and news-test 2014, respectively.

For Finnish-English translation, the training data came from the WMT 2015 shared translation task, which consists of 1.93M sentence pairs with 52.7M English words and 37.8M Finnish words. In this experiment, we used news-dev 2015 as validation set and reported results on news-test 2015.

We evaluated the translations using BLEU [18][2] and TER [23]. Besides, we concatenated all test sets and conducted *boostrap resampling* [12][3] to test the significance in BLEU score differences. For all translation tasks, we employed *Byte Pair Encoding* (BPE) [20] to convert all sentences into subwords, where the source and target sides of the training corpus were combined to learn a joint BPE segmentation with 32000 merge operations. We trained various NMT models with *Adam* [11] SGD for 200,000 steps and selected the optimal model parameters according to the model performance on the validation set. During this procedure, we used the following hyper-parameters: 512 for word embedding dimension, 512 for hidden layer size, 1.0 for gradient norm, 0.3 for dropout rate, and 0.1 for the value of label smoothing. Besides, we used the learning rate decay schedule adopted by Vaswani et al. [27] to adjust the learning rate, and batched sentence pairs with approximately 6000 source and target tokens.

[1] The training set consists of sentence-level parallel corpora LDC2002E18, LDC2003E07, LDC2003E14, news part of LDC2004T08 and document-level parallel corpora LDC2002T01, LDC2004T07, LDC2005T06, LDC2005T10, LDC2009T02, LDC2009T15, LDC2010T03.

[2] We calculate BLEU scores using the `multi-bleu.perl` script.

[3] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/bsbleu.py.

Table 1
Comparison of model parameters and training and decoding speeds (token/second). In our work, we used a single GTX 1080Ti GPU to train different NMT models. In particular, we employed the batch decoding technique to accelerate decoding speed at test time.

SYSTEM	#Para	Training speed	Decoding speed
RNNSearch	54.4M	11,925	816
RNNSearch(R2L)	54.4M	11,925	816
ATNMT	108.8M	11,925	413
NSC(RT)	62.3M	7,116	329
NSC(HS)	58.6M	10,274	377
ABDNMT	78.6M	5,719	524

4.2. Baselines

We compared ABDNMT against the following state-of-the-art NMT models:

- ***RNNSearch***: a re-implementation of the attention-based NMT system [2], which is enhanced by several strategies commonly-used in Transformer [27].
- ***RNNSearch(R2L)***: a variant of RNNSearch that generates translation in a right-to-left manner.
- ***ATNMT***: an attention-based NMT system implemented with two directional decoders [15], which explores the **agreement on target-bidirectional NMT**. When using this model, we first run beam search for forward and backward models independently to generate two k-best lists, which are then combined and re-scored by these two models to find the best candidate. Following Liu et al. [15], we set both beam sizes of two decoders as 10. In particular, we replaced LSTM adopted in [15] with GRU to ensure fair comparison.
- ***NSC(RT)***: a variant of **neural system combination** framework proposed by Zhou et al. [32]. It first uses an attentional NMT model consisting of one encoder and one backward decoder to produce the 1-best reverse translation. Then, it introduces a reverse translation encoder to embed the 1-best reverse translation. Ultimately, in a way similar to the multi-source NMT model [33], another attentional NMT model generates the final output from its standard encoder and the above reverse translation encoder. Different from ours, this model is not an end-to-end one. Moreover, it considers **the embedded hidden states of the reverse translation**, while our model focuses on the hidden states produced by the backward decoder.
- ***NSC(HS)***: it is similar to NSC(RT), however, instead of using a reverse translation encoder to model the 1-best reverse translation, it directly considers **the reverse hidden states produced by the backward decoder**. Likewise, it can not be trained in an end-to-end fashion.

The beam sizes of all above-mentioned models were set as 10, while those of the backward and forward decoders of our model were set as 1 and 10, respectively.

To ensure fair comparison, we equipped all RNNSearch style models with a four-layer RNN decoder. Particularly, as described previously, we employed several strategies to enhance the above-mentioned NMT models: label smoothing, multi-head attention and position-wise feed-forward networks, all of which have been proven effective in Transformer [27]. Besides, we applied layer normalization and dropout to outputs of GRU layers. Dropout was also applied to the source and target word embeddings. We added positional encodings to the word embeddings fed into decoder, and directly tied the parameters of pre-softmax linear transformation to the target word embeddings.

4.3. Results on Chinese-English translation

Model Parameters and Speeds. We first compared the differences of model parameters and efficiency between different NMT models. From Table 1, we observe that compared with most existing models, our models have more parameters and its training speed is relatively slow.

Effect of λ. The hyper-parameter λ reflects the effect of the second term in the training objective. Thus, we first investigated the impact of the hyper-parameter λ (see Eq. (14)) on the validation set. To this end, we gradually varied λ from 0.5 to 1.0 with an increment of 0.1 in each step. Besides, we also compared with a variant of our model, which does not perform greedy beam search but directly uses **reverse reference** during the model training. To clearly depict the experimental results, we referred it as **ABDNMT(RR)**. As shown in Fig. 2, we find that no matter which λ, our model consistently outperforms its variant ABDNMT(RR). This result strongly demonstrates the advantage of the consistency between model training and testing. Particularly, our model achieves the best performance when λ is set as 0.6. Due to this reason, we directly set λ as 0.6 in the following Chinese-English experiments.

The experimental results evaluated by different metrics on Chinese-English translation are shown in Table 2. Particularly, we also displayed the performance of Transformer reported in [30], so as to prove that our models are competitive. Specifically, ABDNMT remarkably outperforms RNNSearch, RNNSearch(R2L), ATNMT, NSC(RT) and NSC(HS) by 1.44/-3.02, 2.21/-3.76,

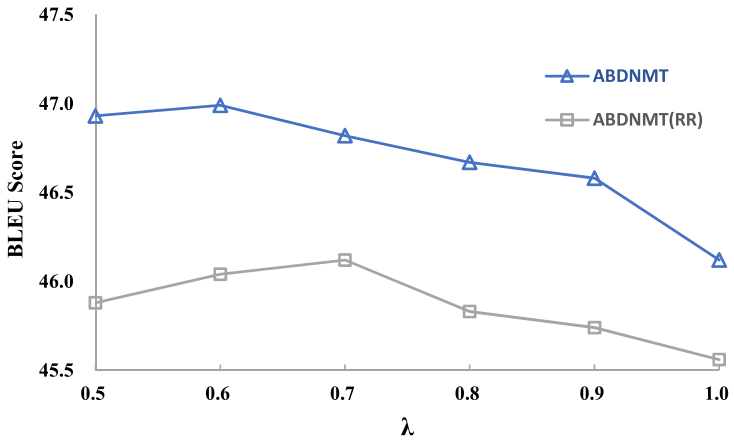

Fig. 2. Experimental results of the NIST Chinese-English translation task on the validation set using different λs.

Table 2
Evaluation of the NIST Chinese-English translation task using case-insensitive BLEU scores (λ=0.6). Here we also cited the performance of Transformer on the same dataset, reported by [30]. ↓/⇓: significantly worse than ABDNMT ($p<0.05/0.01$).

METRIC	SYSTEM	MT02	MT03	MT04	MT05	AVE.
BLEU	Transformer [30]	48.63	47.54	47.79	48.34	48.07
	RNNSearch	47.24	47.05	47.79	46.37	47.11⇓
	RNNSearch(R2L)	46.27	46.53	47.14	45.42	46.34⇓
	ATNMT	47.76	47.63	48.15	46.82	47.59⇓
	NSC(RT)	47.39	47.64	48.19	46.75	47.49⇓
	NSC(HS)	47.55	47.77	48.35	46.94	47.65⇓
	ABDNMT	**47.96**	**48.45**	**49.66**	**48.13**	**48.55**
TER	RNNSearch	51.31	52.11	51.21	52.79	51.86
	RNNSearch(R2L)	51.85	52.73	52.26	53.54	52.60
	ATNMT	49.83	50.26	49.39	51.25	50.18
	NSC(RT)	50.74	50.12	49.17	50.74	50.19
	NSC(HS)	50.25	49.63	48.95	50.88	49.93
	ABDNMT	**49.49**	**48.72**	**47.97**	**49.18**	**48.84**

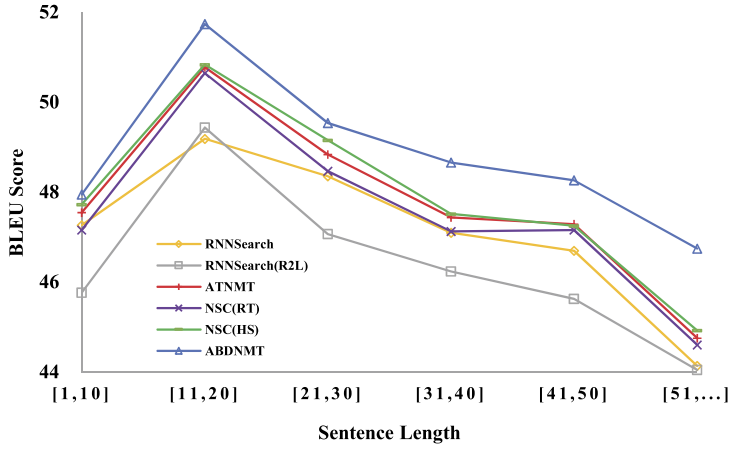

Fig. 3. BLEU scores on different translation groups divided according to source sentence lengths.

0.96/-1.34, 1.06/-1.35, and 0.90/-1.09 BLEU/TER points, respectively. Even when compared with the Transformer reported in [30], ABDNMT still shows better performance.

Furthermore, we divided our test sets into several groups according to the length of source sentences, and compared the system performances in each group. Figs. 3 and 4 illustrates the BLEU and TER scores on these groups of test sets, respec-

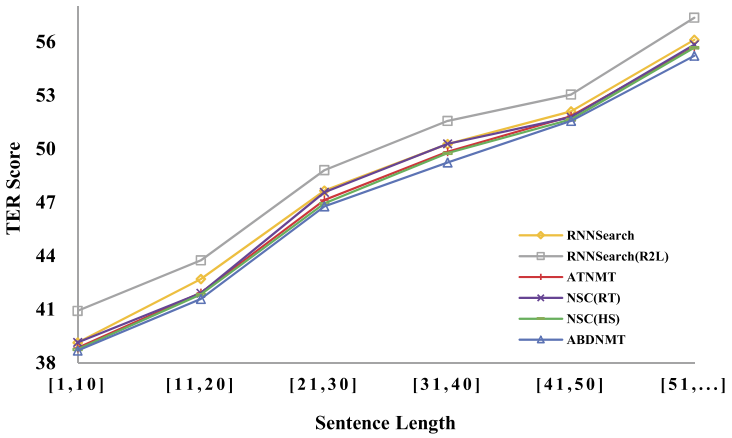

Fig. 4. TER scores on different translation groups divided according to source sentence lengths.

tively. We notice that although the performances of all systems drop with the increase of the length of source sentences, ABDNMT achieves the best performance in most groups. These results further demonstrate the effectiveness of our model.

Finally, we reach the following conclusions based on the above experimental results:

(1) Compared with RNNSearch and RNNSearch(R2L), our model has much better performance, which verifies our assumption that the backward decoder can be used as supplement to the forward decoder in target-side context modeling, and consequently, the simultaneous exploration of bidirectional decoders can lead to better translations.

(2) On all test sets, our model performs better than ATNMT. These results show that joint modeling with attention on reverse hidden states behaves better than k-best hypotheses rescoring [14] in exploiting reverse target-side contexts. This is because our model has two advantages over k-best hypotheses rescoring: On the one hand, the reverse hidden states encode richer target-side contexts than single translation; On the other hand, our model is able to refine translation at a more fine-grained level via the attention mechanism.

(3) The result that NSC(HS) outperforms NSC(RT) reveals that the reverse hidden state representations generated by the backward decoder have distinct advantage in overcoming the problem of data sparsity. Besides, the fact that our model behaves better than NSC(HS) confirms our intuition about the capability of joint model in alleviating the error propagation when encoding target-side contexts.

4.4. Case study

Next, we compared the 1-best translations produced by different models, so as to better understand how our model outperforms others.

Tables 3 and 4 show two Chinese-English translation examples, respectively. We find that RNNSearch tends to produce the translations with good prefixes, while RNNSearch(R2L) generates the translations with desirable suffixes. Although there exist various models with bidirectional decoding that could exploit bidirectional contexts, most of them are unable to precisely translate the whole sentence, while our model is currently the only one capable of producing high quality translations in both examples.

4.5. Analysis

We conducted two groups of experiments to explore which factors enable our model outperforms others.

In the first group of experiments, in order to exclude that the improvement of our proposed model is merely an effect of the increased parameter and computation budget, we compared a variant **ABDNMT**$^-$ of our model, all of which embedding sizes are set as 400, to both RNNSearch and RNNSearch(R2L). By doing so, the parameter size of this model is slightly less than that of RNNSearch.

Table 5 reports the experimental results. Although with less parameters, ABDNMT$^-$ is still obviously superior to both RNNSearch and RNNSearch(R2L). Consequently, we confirm that advantages of our model are not simply due to the increased parameters.

In the second group of experiments, to clearly identify whether the improvement of our proposed model over the conventional RNNSearch comes from two-pass translations, or the reverse target-side context, or a mix of both, we compared our model with the following models, in addition to RNNSearch and RNNSearch(R2L):

- ABDNMT(λ=0). It is a variant of our model, where the backward decoder is not directly trained with a reference but only receives the gradients from the forward decoder;

Table 3
The first Chinese-English translation example of different systems. Texts highlighted in wavy lines are incorrectly translated. In particular, <***UNDER TRAN***> indicates that the under-translation phenomenon occurs in the selected translation. Please note that the translations produced by RNNSearch and RNNSearch(R2L) are generally complementary to each other, and the translation generated by our model is the most accurate and complete.

SRC	冷战结束后，统一后的德国更为强大，西欧力量格局失衡，法国疑@@心增大，致使“德法轴@@心”曾一度运转失灵，使英@@法再度接近，而其意图仍是牵制德国。
REF	*after the cold war is over , the unified germany is stronger ; the power frame in west europe is out of balance . france is more dubious and as a result the " german - france axis " was once out of operation . therefore britain and france come closer with an intention still to contain germany .*
L2R	*after the end of the cold war , the unified germany has become even more powerful . and the pattern of western european forces was unbalanced , and france 's doubts increased . as a result , the " germany - germany axis " failed to operate at one time , making the english law closer to that of germany* <**UNDER TRAN**>*, and its intention* <**UNDER TRAN**> *to contain germany .*
R2L	*after the end of the cold war , after the end of the cold war , germany became more powerful after the end of the cold war , the structure of the western forces is unbalanced , and the suspicion of france has increased . as a result , the " germany - france axis " stops functioning for a time , making the Britain and france get closer, and its intention is still to contain germany .*
ATNMT	*after the end of the cold war , after the end of the cold war , germany became more powerful after the end of the cold war , and the pattern of western european forces was unbalanced . as a result , the " axis of germany " once failed to operate , making the english law closer to it , and its intention* <**UNDER TRAN**> *to contain germany .*
NSC(RT)	*after the end of the cold war , the unified germany became even more powerful . the western* <**UNDER TRAN**> *power structure was out of balance , and the french suspicions increased . as a result , the " axis of france " once operated poorly , causing the french side to get closer to it , and its intention was still to contain germany .*
NSC(HS)	*after the end of the cold war , the unified germany became even more powerful . the western european power structure was unbalanced , and france 's doubts increased . as a result , the " axis of france " once failed to fun, causing the french side to get closer to it , and its intention was still to contain germany .*
ABDNMT	*after the end of the cold war , germany became more powerful . the structure of western european forces became unbalanced , and france 's doubts increased . as a result , the " germany france axis " failed to function for a time , making the england and france closer , and its intention was still to contain germany .*

Table 4
The second Chinese-English translation example of different systems.

SRC	「雅加达邮报」与国营「安@@塔@@拉」新闻通讯社今天纷纷报导，雅加达当局以破坏美观的理由，否绝@@了这项要求以围墙取代目前将使馆和大使官邸与外界隔@@开的篱@@笆的计画。
REF	*the jakarta post " and the state - owned " antala " news agency both reported today that jakarta authorities rejected this proposal to build a wall to replace the fence that now separates the embassy and the residences from the outside on grounds that it is aesthetically unacceptable .*
L2R	*the jakarta post and state - run antara news agency reported today that the jakarta authorities have ignored the request for a wall to replace the residence of the embassy and ambassador 's residence with the exception of the outside world* <**UNDER TRAN**> *.*
R2L	*the jakarta post and the state - run anatolia news agency reported today that there was no reason for the jakarta authorities to use the walls to replace the current plan to separate the embassy and ambassador 's residence from the outside world for reasons of beauty .*
ATNMT	*the jakarta post and the state - run anatolia news agency reported today that the jakarta authorities have ignored the jakarta authorities ' plan to replace the existing* <**UNDER TRAN**> *residence of the embassy and ambassador 's residence with the outside world* <**UNDER TRAN**> *.*
NSC(RT)	*the jakarta post and the state - run antara news agency reported today that the jakarta authorities had no reason to use the wall to replace the current plan to separate the embassy from the embassy and the outside world on the grounds that the jakarta authorities were trying to destroy their beauty .*
NSC(HS)	*the jakarta post and the state run antara news agency reported today that for a beautiful reason , the jakarta authorities had to be replaced by a wall to replace the current plan of separating the embassy and ambassador 's residence with the outside world .*
ABDNMT	*the jakarta post and the state - run antara news agency reported today that the jakarta authorities rejected to use walls to replace the current plan to separate the embassy and the ambassador 's residence from the outside world for reasons of damaging beauty .*

Table 5
Case-insensitive BLEU score of different models using similar sizes of parameters (λ=0.6). $\downarrow/\Downarrow$: significantly worse than ABDNMT$^-$ ($p<0.05/0.01$).

SYSTEM	#Para	MT02	MT03	MT04	MT05	Ave.
RNNSearch	54.4M	47.24	47.05	47.79	46.37	47.11$^{\Downarrow}$
RNNSearch(R2L)	54.4M	46.27	46.53	47.14	45.42	46.34$^{\Downarrow}$
ABDNMT$^-$	53.2M	47.70	48.13	48.95	47.47	48.06

Table 6
Evaluation of the NIST Chinese-English translation task using case-insensitive BLEU scores. $\downarrow/\Downarrow$: significantly worse than ABDNMT ($p<0.05/0.01$).

SYSTEM	MT02	MT03	MT04	MT05	Ave.
RNNSearch	47.24	47.05	47.79	46.37	47.11$^{\Downarrow}$
RNNSearch(R2L)	46.27	46.53	47.14	45.42	46.34$^{\Downarrow}$
ABDNMT(λ=0)	47.60	47.36	48.38	46.90	47.56$^{\Downarrow}$
ABDNMT(FF)	47.68	48.21	48.54	47.62	48.01$^{\downarrow}$
ABDNMT	47.96	48.45	49.66	48.13	48.55

Table 7
Case-insensitive BLEU scores with length penalty of the left and right half of ABDNMT(FF) and ABDNMT. $\downarrow/\Downarrow$: significantly worse than ABDNMT ($p<0.05/0.01$).

SYSTEM	Half	MT02	MT03	MT04	MT05	Ave.
ABDNMT(FF)	Left	51.66	53.73	53.39	51.59	52.59
	Right	45.27	43.92	46.11	45.20	45.13$^{\Downarrow}$
ABDNMT	Left	51.57	53.15	53.09	51.52	52.33
	Right	46.05	45.33	48.53	46.68	46.65

Table 8
Evaluation of the WMT English-German translation task using case-sensitive BLEU and TER scores (λ=0.6). We directly cited the performance of Transformer reported in [1]. $\downarrow/\Downarrow$: significantly worse than ABDNMT ($p<0.05/0.01$).

SYSTEM	BLEU	TER
Transformer [1]	27.30	–
RNNSearch	26.04$^{\Downarrow}$	57.24
RNNSearch(R2L)	25.20$^{\Downarrow}$	59.02
ATNMT	26.58$^{\Downarrow}$	56.81
NSC(RT)	26.39$^{\Downarrow}$	57.15
NSC(HS)	26.54$^{\downarrow}$	56.69
ABDNMT	**27.15**	**56.23**

- ABDNMT(FF). As a different variant of our model, its backward decoder is replaced by another forward decoder. In other words, its two-pass translations are both implemented in left-to-right way.

From Table 6, we obtain the following two observations: First, both ABDNMT(λ=0) and ABDNMT(FF) achieve better performance than the conventional RNNSearch and RNNSearch(R2L), suggesting two-pass translations are indeed beneficial to NMT. Second, our proposed model significantly surpasses both ABDNMT(λ=0) and ABDNMT(FF). Further, we split each generated translation of ABDNMT(FF) or ABDNMT equally into two halves, and then evaluated the translation qualities of these two halves separately in Table 7. We can find that both models have similar performance on generating the left part of translations, while ABDNMT is obviously better than ABDNMT(FF) when producing the right part of translations. These results strongly show that the reverse target-side context is indeed complementary to the left-side context utilized by NMT models.

4.6. Results on English-German translation

Also, to enhance the credibility of our experiments, we provided the performance of ***Transformer*** reported in [1]. According to the performance of our model on the validation set, we determined the optimal λ as 0.6.

Table 8 shows the results on English-German translation. Our model still significantly outperforms others, even is comparable to the Transformer reported in [1]. It should be noted that the BLEU score gaps between our model and the others on

Table 9
Evaluation of the WMT Finnish-English translation task using case-sensitive BLEU and TER scores (λ=0.7). We also listed the performance of CharNMT reported in [3]. CharNMT is a character-level deep NMT model. $\downarrow/\Downarrow$: significantly worse than ABDNMT ($p<0.05/0.01$).

SYSTEM	BLEU	TER
CharNMT	19.30	–
RNNSearch	$18.41^{\Downarrow}$	66.33
RNNSearch(R2L)	$18.03^{\Downarrow}$	67.31
ATNMT	$18.89^{\downarrow}$	66.01
NSC(RT)	$18.72^{\Downarrow}$	65.86
NSC(HS)	$18.85^{\downarrow}$	65.43
ABDNMT	**19.64**	**65.06**

Table 10
Evaluation of the NIST Chinese-English translation task using case-insensitive BLEU and TER scores (λ=0.6). $\downarrow/\Downarrow$: significantly worse than ABDTransformer ($p<0.05/0.01$).

METRIC	SYSTEM	MT02	MT03	MT04	MT05	AVE.
BLEU	Transformer [30]	48.63	47.54	47.79	48.34	48.07
	Transformer	48.27	48.12	49.25	49.00	$48.66^{\Downarrow}$
	ABDTransformer	**49.81**	**49.94**	**51.09**	**50.05**	**50.22**
TER	Transformer	48.75	49.25	48.67	48.24	48.73
	ABDTransformer	**47.31**	**47.51**	**47.07**	**47.05**	**47.24**

English-German translation are much smaller than those on Chinese-English translation. The underlying reasons lie in the following two aspects, which have also been mentioned in [22]. First, the Chinese-English datasets contain four reference translations for each sentence, while the English-German dataset only have single reference. Second, compared with German, Chinese is more indistinctly related to English, making the advantage of utilizing target-side contexts more remarkable in Chinese-English translation.

4.7. Results on Finnish-English translation

Similarly, we first set the optimal λ as 0.7, on the basis of the performance of our model on the validation set. In this group of experiments, we provided the performance of ***CharNMT*** [3] on the same dataset to enhance the persuasiveness of our experiments.

The experimental results on Finnish-English translation are reported in Table 9. In general, our model achieves the highest BLEU and the lowest TER scores again, with significant improvements over those of all contrast systems.

4.8. Extension of our method into transformer

Recently, due to advantages on long-range dependency modeling and flexibility in parallel computation, Transformer [27] has become the state-of-the-art model in the community of machine translation. Here we extend our method into Transformer, so as to verify the generality of our approach in different NMT frameworks. To this end, we also equip the conventional Transformer with a backward decoder, where its hidden states via greedy search are then considered by the forward decoder to produce the final translation. Likewise, this enhanced Transformer can be trained in an end-to-end way using the objective function shown in Eq. (14).

We implemented Transformer and its enhanced variant **ABDTransformer** based on the open-source *THUMT*.[4] We used the same datasets and model configuration including λ as those of experiments related to RNNSearch. The only difference is that we batched sentence pairs with approximately 24000 source and target tokens. In this way, we can ensure the fair comparison with previous work [30].

Table 10, 11 and 12 report the final experimental results. We can see that no matter for which language pairs, ABDTransformer significantly outperforms the standard Transformer. Therefore, we believe that our framework is effective and general to both RNNSearch and Transformer.

[4] https://github.com/thumt/THUMT.

Table 11
Evaluation of the WMT English-German translation task using case-sensitive BLEU and TER scores (λ=0.6). $\downarrow/\Downarrow$: significantly worse than ABDTransformer ($p<0.05/0.01$).

SYSTEM	BLEU	TER
Transformer [1]	27.30	–
Transformer	27.83$\Downarrow$	56.15
ABDTransformer	**29.59**	**53.66**

Table 12
Evaluation of the WMT Finnish-English translation task using case-sensitive BLEU and TER scores (λ=0.7). $\downarrow/\Downarrow$: significantly worse than ABDTransformer ($p<0.05/0.01$).

SYSTEM	BLEU	TER
CharNMT	19.30	–
Transformer	20.03$\Downarrow$	64.31
ABDTransformer	**21.32**	**62.98**

5. Conclusions and future work

In this paper, we have extended the conventional attentional encoder-decoder NMT model by introducing a backward decoder. In our model, the first step of decoding is to apply the backward decoder to generate hidden states encoding reverse target-side contexts. Then, via two different attention mechanisms, the forward decoder simultaneously considers two individual hidden state sequences produced by the encoder and the backward decoder to generate the final translation. In contrast to the previous models, ours is an end-to-end NMT model that fully utilizes reverse target-side contexts for translation. Experimental results on several different translation directions indicate the effectiveness of our model.

Our model is generally applicable to other models with RNN-based decoder. Therefore, the effectiveness of our approach on other NMT models [24,25] and tasks related to RNN-based decoder modeling, such as image captioning, will be investigated in future research. Moreover, in our work, the attention mechanisms acting on decoders in two directions are independent from each other. However, intuitively, these two mechanisms should be closely associated with each other. Therefore, we are interested in exploring better attention mechanism combination to further refine our model.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No. 2017YFB0202204), the National Natural Science Foundation of China (Nos. 61672440, 61761166008, and 61432013), Beijing Advanced Innovation Center for Language Resources (No. TYR17002), the Fundamental Research Funds for the Central Universities (No. ZK1024) and the Scientific Research Project of National Language Committee of China (No. YB135-49).

References

[1] K. Ahmed, N.S. Keskar, R. Socher, Weighted transformer network for machine translation, CoRR, arXiv:1711.02132, 2017.
[2] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proc. of ICLR2015, 2015.
[3] C. Cherry, G. Foster, A. Bapna, O. Firat, W. Macherey, Revisiting character-based neural machine translation with capacity and compression, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 – November 4, 2018, 2018, pp. 4295–4305.
[4] D. Chiang, Hierarchical phrase-based translation, Comput. Linguist. 33 (2007) 201–228.
[5] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder–decoder for statistical machine translation, in: Proc. of EMNLP2014, 2014, pp. 1724–1734.
[6] A. Finch, E. Sumita, Bidirectional phrase-based statistical machine translation, in: Proc. of EMNLP2009, 2009, pp. 1124–1132.
[7] C.D.V. Hoang, G. Haffari, T. Cohn, Towards decoding as continuous optimisation in neural machine translation, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, 9–11 September, 2017, 2017, pp. 146–156.
[8] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. (1997) 1735–1780.
[9] M. Junczys-Dowmunt, R. Grundkiewicz, An exploration of neural sequence-to-sequence architectures for automatic post-editing, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing, Volume 1: Long Papers, IJCNLP 2017, Taipei, Taiwan, November 27 – December 1, 2017, 2017, pp. 120–129.

[10] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models, in: Proc. of EMNLP2013, 2013, pp. 1700–1709.
[11] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May, 2015, 2015, Conference Track Proceedings.
[12] P. Koehn, Statistical significance tests for machine translation evaluation, in: Proc. of EMNLP2004, 2004, pp. 388–395.
[13] P. Koehn, F.J. Och, D. Marcu, Statistical phrase-based translation, in: Proc. of NAACL2003, 2003, pp. 48–54.
[14] L. Liu, A. Finch, M. Utiyama, E. Sumita, Agreement on target-bidirectional lstms for sequence-to-sequence learning, in: Proc. of AAAI2016, 2016, pp. 2630–2637.
[15] L. Liu, M. Utiyama, A. Finch, E. Sumita, Agreement on target-bidirectional neural machine translation, in: Proc. of NAACL2016, 2016, pp. 411–416.
[16] J. Niehues, E. Cho, T.L. Ha, A. Waibel, Pre-translation for neural machine translation, in: Proc. of COLING2016, 2016, pp. 1828–1836.
[17] S. Pal, S.K. Naskar, M. Vela, Q. Liu, J. van Genabith, Neural automatic post-editing using prior alignment and reranking, in: Proc. of EACL2017, 2017, pp. 349–355.
[18] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proc. of ACL2002, 2002, pp. 311–318.
[19] R. Sennrich, B. Haddow, A. Birch, Edinburgh neural machine translation systems for WMT 16, in: Proceedings of the First Conference on Machine Translation, WMT 2016, Colocated with ACL 2016, 11–12 August, Berlin, Germany, 2016.
[20] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proc. of ACL2016, 2016, pp. 1715–1725.
[21] I.V. Serban, A. Sordoni, Y. Bengio, A. Courville, J. Pineau, Twin networks: matching the future for sequence generation, in: Proc. of ICLR2018, 2018.
[22] S. Shen, Y. Cheng, Z. He, W. He, H. Wu, M. Sun, Y. Liu, Minimum risk training for neural machine translation, in: Proc. of ACL2016, 2016, pp. 1683–1692.
[23] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proc. of AMTA2006, 2006.
[24] J. Su, S. Wu, D. Xiong, Y. Lu, X. Han, B. Zhang, Variational recurrent neural machine translation, in: Proc. of AAAI2018, 2018, pp. 5488–5495.
[25] J. Su, J. Zeng, D. Xiong, Y. Liu, M. Wang, J. Xie, A hierarchy-to-sequence attentional neural machine translation model, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (2018) 623–632.
[26] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Proc. of NIPS2014, 2014, pp. 3104–3112.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017, pp. 6000–6010.
[28] T. Watanabe, E. Sumita, Bidirectional decoding for statistical machine translation, in: Proc. of COLING 2002, 2002, pp. 1200–1208.
[29] H. Zhang, K. Toutanova, C. Quirk, J. Gao, Beyond left-to-right: multiple decomposition structures for smt, in: Proc. of NAACL2013, 2013, pp. 12–21.
[30] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, M. Zhang, Y. Liu, Improving the transformer translation model with document-level context, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 533–542.
[31] X. Zhang, J. Su, q. Que, L. Yang, J. Rongrong, W. Hongji, Asynchronous bidirectional decoding for neural machine translation, in: Proc. of AAAI2018, 2018, pp. 5698–5705.
[32] L. Zhou, W. Hu, J. Zhang, C. Zong, Neural system combination for machine translation, in: Proc. of ACL2017, 2017, pp. 378–384.
[33] B. Zoph, K. Knight, Multi-source neural translation, in: Proc. of NAACL2016, 2016, pp. 30–34.