Synchronous Interactive Decoding for Multilingual Neural Machine Translation

Hao He,[1,2] **Qian Wang,**[1,2] **Zhipeng Yu,**[3] **Yang Zhao,**[1,2] **Jiajun Zhang,**[1,2] **Chengqing Zong**[1,2]
[1]National Laboratory of Pattern Recognition, CASIA, Beijing 100190, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Beijing Fanyu Technology Co., Ltd, Beijing 100083, China
hao.he@ia.ac.cn, qian.wang@nlpr.ia.ac.cn, yusade123@gmail.com,
zhaoyang2015@ia.ac.cn, {jjzhang, cqzong}@nlpr.ia.ac.cn

Abstract

To simultaneously translate a source language into multiple different target languages is one of the most common scenarios of multilingual translation. However, existing methods cannot make full use of translation model information during decoding, such as intra-lingual and inter-lingual future information, and therefore may suffer from some issues like the unbalanced outputs. In this paper, we present a new approach for synchronous interactive multilingual neural machine translation (SimNMT), which predicts each target language output simultaneously and interactively using historical and future information of all target languages. Specifically, we first propose a synchronous cross-interactive decoder in which generation of each target output does not only depend on its generated sequences, but also relies on its future information, as well as history and future contexts of other target languages. Then, we present a new interactive multilingual beam search algorithm that enables synchronous interactive decoding of all target languages in a single model. We take two target languages as an example to illustrate and evaluate the proposed SimNMT model on IWSLT datasets. The experimental results demonstrate that our method achieves significant improvements over several advanced NMT and MNMT models.

Introduction

Neural machine translation (NMT) has greatly improved the translation quality (Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2015; Gehring et al. 2017; Vaswani et al. 2017) and promoted the studies on multilingual translation. Due to the powerful end-to-end modeling capability based on the encoder-decoder framework, it is possible to handle multiple language pairs in a single model. As the deployment cost among multiple languages pairs is significantly reduced, the single model-based approach becomes a promising paradigm in multilingual NMT (MNMT).

Training a single model with multiple language pairs can leverage the complementary information of different languages (Johnson et al. 2017), such as enabling zero-resource translations, or improving the quality of low-resource translations. However, existing methods cannot make full use of the information in the model during decoding. Some work (Johnson et al. 2017; Wang et al. 2018) support multilingual translation with single model, but the translation for each sentence in a batch is independent. Multi-target translation (Dong et al. 2015) supports to translate a source sentence into several different target languages through a model with one shared encoder and several different decoders. Although employing multiple decoders, the model can still handle only one language pair at each moment during decoding, which brings two problems: (i) the decoding process cannot use complementary information among different languages; (ii) for only depending on historical information without using future information, it suffers from the issue of unbalanced target language generations, i.e., the prefixes of sentences are better predicted than the suffixes (Liu et al. 2016).

Several studies have explored these two issues. To exploit the multilingual complementary information, Wang et al. 2019 synchronously translates a sentence into two different target languages, and allows the generated sequences to attend to another language's ongoing generation, to improve translation quality. However, since only historical information is adopted in the translation, it still faces the unbalanced output problem. Some studies (Liu et al. 2016; Zhang et al. 2018; Zhou, Zhang, and Zong 2019; Zhou et al. 2019; Zhang et al. 2020) alleviate this problem by introducing bidirectional decoding which provides both historical and future information during decoding. However, these works cannot support bidirectional decoding for multiple languages in a single decoder. Actually, the multilingual conversation is quite a common scenario where one sentence needs to be simultaneously translated into multiple other languages (e.g., international group chat, conversation and meeting, etc.). Therefore, it is a meaningful and promising direction to design a synchronous interactive multilingual NMT model that enables the historical and future information of different target languages to interact with each other to improve the translation performance.

For example (in Figure 1), the traditional NMT model translates an English sentence into a Chinese sentence from-left-to-right (L2R), which only depends on the historical information that has been generated (only blue box). However, for multi-target MNMT, all the forward (L2R) and backward (from-right-to-left, R2L) sequences of different target languages share the same semantics. Therefore, at each step

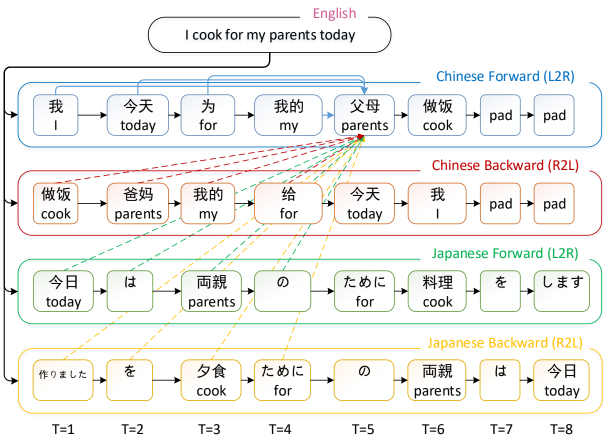


Figure 1: This illustration gives a simple example of translation decoding which uses 4 types of information (historical and future information of intra-language and inter-languages) to predict current token.

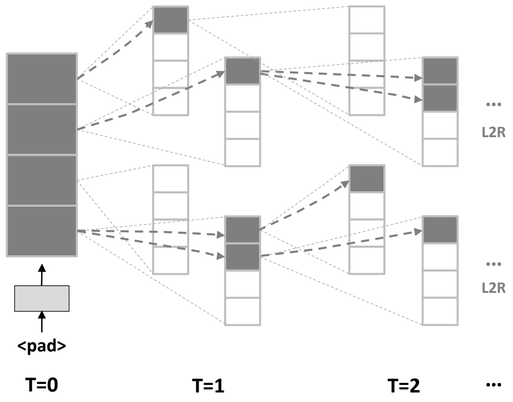


Figure 2: Illustration of standard beam search algorithm with beam size 4, where the dark blocks denote the alive hypotheses (the ongoing expansions).

of decoding, different target languages with different translation directions can provide more information to predict a token. Specifically, when generating a target sequence, the prediction of each token can rely on the 4 types of information (4 color boxes), that is, the intra-lingual historical and future information, and inter-lingual historical and future information. For the sake of brevity (as shown in Figure 1), we only illustrate the prediction of token "parents" in Chinese forward sequence. For the other 3-way generations, 4 kinds of information can also be used for decoding.

Accordingly, in this work, we propose a novel framework (SimNMT) which simultaneously generates multiple target sentences by interactively using intra-lingual and inter-lingual information. Specifically, we first present a synchronous cross-interactive decoder, in which the generation of each target language does not only depend on its previously generated sequences, but also its future context (intra-lingual information), as well as the historical and future information of other target languages (inter-lingual information). Then, we propose a new interactive multilingual beam search algorithm that enables synchronous interactive decoding of all target languages in a single model.

The major contributions of this paper are as follows:

(1) We propose a synchronous cross-interactive attention, which interacts historical and future information of intra-language and inter-languages, to improve multiple target language translation qualities. To our best knowledge, this is the first work to explore the effectiveness of a single NMT model with these 4 types of information interaction.

(2) To cooperate with the proposed decoder, we present a synchronous interactive multilingual decoding algorithm. It maintains both forward and backward hypotheses for each target language during translation, and all hypotheses perform information interaction in each step of decoding.

(3) The experiments have demonstrated that the proposed synchronous interactive multilingual NMT (SimNMT) shows significant improvements over both the previous competitive NMT and MNMT models.

Background

In this paper, we build our model based on Transformer (Vaswani et al. 2017) with encoder-decoder framework. The encoder first encodes an input sequence of symbols x = (x_1, x_2, ..., x_n) to a sequence of continues representations z = (z_1, z_2, ..., z_n), then the decoder generates an output sequence y = (y_1, y_2, ..., y_m) token-by-token.

Multi-Head Attention allows the model to attend to information from different representation subspaces at different positions. The calculation is based on queries Q, keys K, and values V. For multi-head self-attention in encoder or decoder of standard Transformer, all of the Q, K, V are output hidden-state matrices of the previous layer. For multi-head inter-attention in decoder, Q is the hidden state matrix of the previous decoder layer, and K-V pairs are the outputs (z_1, z_2, ..., z_n) of the encoder.

To calculate multi-head attention, it first needs to obtain h different representations of (Q_i, K_i, V_i). Each attention head i projects the hidden-state matrix into independent query, key, and value representations $Q_i = QW_i^Q$, $K_i = KW_i^K$, $V_i = VW_i^V$, respectively. Then, we perform scaled dot-product attention for each representation, concatenate them and send the concatenation into a feed-forward layer.

$$\mathrm{MhAtt}(Q, K, V) = \mathrm{Concat}_i(head_i)W^O \tag{1}$$

$$head_i = \mathrm{Attention}(Q_i, K_i, V_i) \tag{2}$$

where W_i^Q, W_i^K, W_i^V and W^O are parameter projection matrices.

Scaled Dot-Product Attention: We first multiply query Q_i by key K_i to obtain an attention weight matrix, which is then multiplied by value V_i for each token to obtain the self-attention representation. The scaled dot-product attention is calculated by a query Q, a key K, and a value V:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

Method	Target Language(s) Information Utilization				
	Intra-lingual History	Intra-lingual Future	Inter-lingual History	Inter-lingual Future	Additional Corpus
Transformer	✓				
Transformer (+pseudo)	✓				✓
GNMT-Multi	✓		✓		✓
SB-NMT	✓	✓			✓
SyncTrans	✓		◇	◇	✓
SimNMT	✓	✓	✓	✓	✓

Table 1: Translation Information Dependencies.

where d_k is the dimension of the key.

Standard Beam Search: We usually adopt beam search (greedy search if beam size = 1) to pick the best translation y for the input x via formula $\hat{y}$ = softmax$_y P(y|x)$. Beam size N is used to control the search space by expanding only the top-N hypotheses each round. As shown in Figure 2, the 4 dark blocks represent the 4 best token expansions of the last state, and these token expansions are sorted top-to-bottom by probabilities. The translation y is generated token-by-token, and a complete hypothesis is defined as a hypothesis which outputs end-of-sentence (EOS) symbol.

Our Approach

As discussed in Section 1, on one hand, historical and future (intra-lingual) information of a target language can be used to improve translation quality. On the other hand, when a sentence is translated into different target languages, the different (inter-lingual) outputs can be complementary to each other. Therefore, it is reasonable to improve translation performance by integrating all of above information.

In this section, we will introduce the approach of intra-lingual and inter-lingual attention synchronous interactive multilingual NMT (SimNMT). Our goal is to design a decoder to translate one source language into different target languages, in which generation of each target output does not only depend on its historical and future information, but also relies on history and future contexts of other target language(s). The core module is synchronous interactive multilingual attention (SimAtt, see 3.1), which replaces the multi-head intra-attention in the decoder of the Transformer model (see 3.2). To cooperate with the proposed decoder in decoding, we present a new synchronous interactive multilingual decoding algorithm, which can decode all target languages in a synchronous interactive way (see 3.3).

From viewpoint of information utilization, Table 1 lists the translation dependencies of several NMT models. Different from the standard Transformer (Vaswani et al. 2017), when translating a certain target language, the other 5 methods all utilize additional data. Particularly, GNMT-Multi (Johnson et al. 2017) and SyncTrans (Wang et al. 2019) contain different target languages information (SyncTrans employs one of forward or backward sequences), while SB-NMT (Zhou, Zhang, and Zong 2019) introduces future information (backward sequences) of one target language. The proposed SimNMT method employs all 4 types of information to for translation.

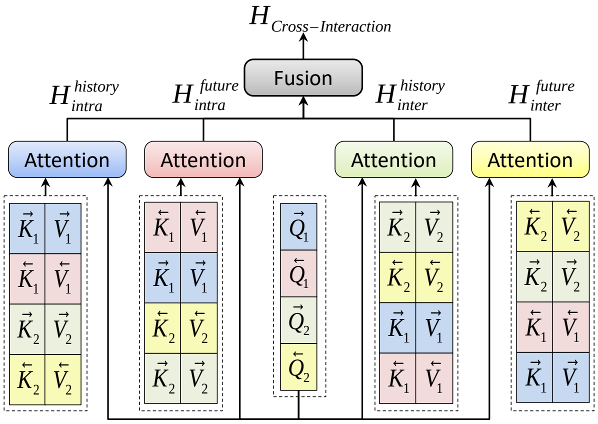


Figure 3: Synchronous interactive multilingual attention (SimAtt) framework. It simultaneously operates on the forward (history) and backward (future) queries Q, keys K, and values V of different target languages.

Compared with the related NMT models, our approach has the following advantages: (1) The multiple target outputs are generated by a single model (one encoder and one decoder), and all target sequences can be processed in parallel. (2) Based on attention synchronous interaction mechanism, the proposed model is an end-to-end joint framework and can optimize multilingual decoding simultaneously. (3) Compared with applying independent beam search to each target output, our decoder is faster and more compact by using one beam search algorithm for all target generations.

Synchronous Interactive Multilingual Attention

We take two target languages as an example to illustrate the proposed synchronous interactive multilingual attention (SimAtt) method. As illustrated in Figure 3, the input consists of queries Q, keys K, and values V which are all concatenated (in different order) by language 1 forward (L1-L2R) states and backward (L1-R2L) states, as well as language 2 forward (L2-L2R) states and backward (L2-R2L) states. Now we only look at the language 1 forward direction, whose forward query $\overrightarrow{Q}_1$ (blue box) simultaneously operates on 4 types of "key-value" pairs to calculate attention, including key and value of: L1-L2R $\overrightarrow{K}_1\overrightarrow{V}_1$ (blue box), L1-R2L $\overleftarrow{K}_1\overleftarrow{V}_1$ (red box), L2-L2R $\overrightarrow{K}_2\overrightarrow{V}_2$ (green box) and L2-R2L $\overleftarrow{K}_2\overleftarrow{V}_2$ (yellow box). It obtains 4 hidden states $\overrightarrow{H}_{1intra}^{history}$, $\overrightarrow{H}_{1intra}^{future}$, $\overrightarrow{H}_{1inter}^{history}$ and $\overrightarrow{H}_{1inter}^{future}$, which attempts to make use of existing information as effectively as possible to help predict the current L1-L2R token during decoding. The hidden state of L1-L2R can be calculated by:

$$\begin{cases} \vec{H}_{1intra}^{history} = \text{Attention}(\vec{Q}_1, \vec{K}_1, \vec{V}_1) \\ \vec{H}_{1intra}^{future} = \text{Attention}(\vec{Q}_1, \overleftarrow{K}_1, \overleftarrow{V}_1) \\ \vec{H}_{1inter}^{history} = \text{Attention}(\vec{Q}_1, \vec{K}_2, \vec{V}_2) \\ \vec{H}_{1inter}^{future} = \text{Attention}(\vec{Q}_1, \overleftarrow{K}_2, \overleftarrow{V}_2) \end{cases} \quad (4)$$

For the other 3 queries $\overleftarrow{Q}_1$, $\overrightarrow{Q}_2$, $\overleftarrow{Q}_2$ (corresponding to L1-R2L, L2-L2R and L2-R2L, respectively), each of

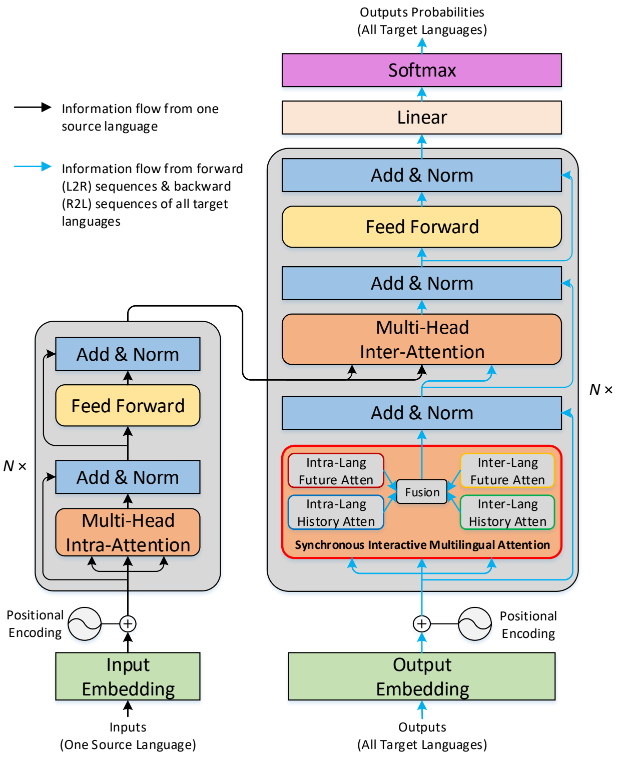


Figure 4: The new Transformer architecture with the proposed synchronous cross-interactive decoder. The input of decoder is concatenation of all target languages' forward (L2R) and backward (R2L) sequence, and these information flow runs in parallel and interacts in synchronous interactive multilingual attention (SimAtt) sub-layer.

them also performs attention calculations with 4 different types of "key-value" pairs, respectively. After that, it will get 4 types of hidden states: $\overrightarrow{H}_{intra}^{history}$, $\overrightarrow{H}_{intra}^{future}$, $\overrightarrow{H}_{inter}^{history}$ and $\overrightarrow{H}_{inter}^{future}$. Overall hidden state $H_{Cross-Interaction}$ is formed by the fusion of above 4 hidden states. The fusion function Fusion($\cdot$) combines these 4 types of hidden states by using linear interpolation, non-linear activation function, or gate mechanism, etc.

Synchronous Interactive MNMT Model

The synchronous interactive multilingual NMT (SimNMT) model applies the proposed SimAtt module to replace the multi-head intra-attention sub-layer in a standard Transformer decoder (as shown in Figure 4), which we name it synchronous cross-interactive decoder. The encoder of our model is identical to standard Transformer.

For both training and inferencing, through SimAtt sub-layer, we allow the forward (L2R) and backward (R2L) information flows of all target languages to interact with each other, which can obtain more information for translation, to mutually enhance the performance of different target language translations corresponding to the same source language sentence. Like the standard Transformer, the SimAtt sub-layer also uses residual connections around it, and followed by layer normalization:

$$h_d^n = \text{Norm}(h_d^{n-1} + \text{SimAtt}(h^{n-1}, h^{n-1}, h^{n-1})) \qquad (5)$$

where n is layer depth, subscript d denotes decoder, h means hidden states. The h_d^{n-1} is equal to $[\overrightarrow{h}_{1d}^{n-1}; \overleftarrow{h}_{1d}^{n-1}; \overrightarrow{h}_{2d}^{n-1}; \overleftarrow{h}_{2d}^{n-1}]$, which is concatenation of forward and backward hidden states of all target languages and can be processed in parallel. After SimAtt module, similar to the standard Transformer, two other sublayers are stacked to attend to the source semantics related to translation:

$$h_e^n = \text{Norm}(h_d^n + \text{MhAtt}(h_d^n, h_e^N, h_e^N)) \qquad (6)$$

$$h^n = \text{Norm}(h_e^n + \text{FFN}(h_e^n)) \qquad (7)$$

where MhAtt means the multi-head attention, h_e^n denotes the top layer hidden state of encoder, and FFN refers feed-forward networks.

Finally, we adopt linear transform and softmax activation functions to obtain the probability of the next 4 tokens based on h^N = $[\overrightarrow{h}_1^N; \overleftarrow{h}_1^N; \overrightarrow{h}_2^N; \overleftarrow{h}_2^N]$:

$$\begin{cases} p(\vec{y}_1^j|\alpha, x, \theta) = \text{Softmax}(\vec{h}_1^N W) \\ p(\overleftarrow{y}_1^j|\alpha, x, \theta) = \text{Softmax}(\overleftarrow{h}_1^N W) \\ p(\vec{y}_2^j|\alpha, x, \theta) = \text{Softmax}(\vec{h}_2^N W) \\ p(\overleftarrow{y}_2^j|\alpha, x, \theta) = \text{Softmax}(\overleftarrow{h}_2^N W) \end{cases} \qquad (8)$$

where α = $(\overrightarrow{y}_1^{<j}, \overleftarrow{y}_1^{<j}, \overrightarrow{y}_2^{<j}, \overleftarrow{y}_2^{<j})$, θ is shared weight for SimNMT inference, and W is the weight matrix.

Synchronous Interactive Decoding for MNMT

Figure 5 illustrates the synchronous interactive multilingual (SIM) beam search process for 2 target languages with beam size 8. For each target language, the proposed SIM beam search approach maintains 2 types of translation hypotheses (forward and backward, i.e., L2R and R2L), which predict and expand token-by-token. When SimNMT predicts next tokens, SimAtt is performed among all hypotheses to attend 4 types information: historical and future information of intra-language and inter-languages.

For 2 target languages with beam size 8, at each time step, each target language translation hypotheses will keep 4-best items (as long as there is at least one alive hypothesis for this target language). The dark colored blocks denote that the hypothesis is still expanding, while light colored blocks indicate the hypotheses have produced EOS symbol. When all sequences produce EOS or exceed the maximum sequence length, the decoding is terminated. For a target language, if a backward sequence score exceeds forward one, the sequence will be reversed before output.

Training

For two target languages, we need to prepare trilingual datasets $\{x^{(z)}, \overrightarrow{y}_1^{(z)}, \overleftarrow{y}_1^{(z)}, \overrightarrow{y}_2^{(z)}, \overleftarrow{y}_2^{(z)}\}_{z=1}^{Z} \in D$, where each target language has two types of training data, forward $\overrightarrow{y}$ and backward $\overleftarrow{y}$ sequences. Note that a pair of forward and backward target training sequences shares a same meaning, but they should not be exactly the same (i.e., the R2L sequences should not be simply reversed from the L2R ones).

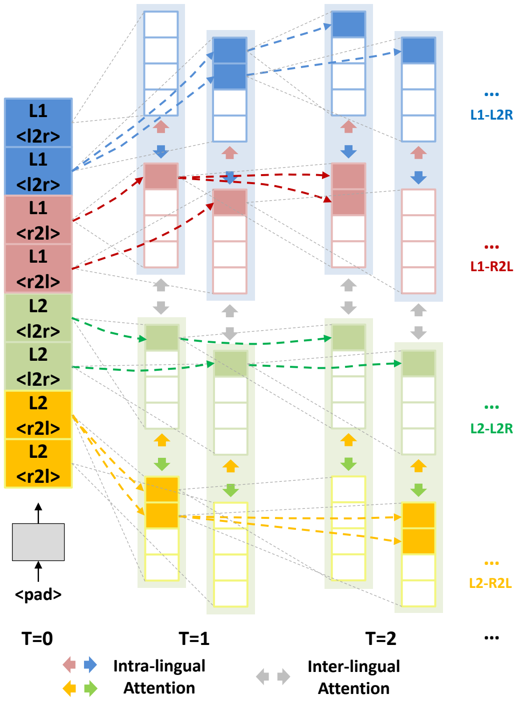


Figure 5: The synchronous interactive multilingual (SIM) beam search with beam size 8. Each target language (e.g., L1 or L2) maintains both forward (L2R) and backward (R2L) translation hypotheses. At each time step (T) of decoding, SimAtt is performed among all hypotheses, which interact intra-lingual and inter-lingual attention with each other. The dark blocks are the alive hypotheses (ongoing expansions), while the light blocks are the sequences with EOS symbols.

The objective function aims to find the model parameters that maximize the log-likelihood over the 4 target sequences:

$$
\begin{aligned}
J(\theta) = \sum_{z=1}^{Z} \sum_{j=1}^{M} \{ & \log p(\vec{y}_1^{j} | \alpha^{(z)}, x^{(z)}, \theta) \\
& + \log p(\overleftarrow{y}_1^{j} | \alpha^{(z)}, x^{(z)}, \theta) \\
& + \log p(\vec{y}_2^{j} | \alpha^{(z)}, x^{(z)}, \theta) \\
& + \log p(\overleftarrow{y}_2^{j} | \alpha^{(z)}, x^{(z)}, \theta) \}
\end{aligned} \tag{9}
$$

where $\alpha = (\overrightarrow{y}_1^{<j}, \overleftarrow{y}_1^{<j}, \overrightarrow{y}_2^{<j}, \overleftarrow{y}_2^{<j})$. Z is sentences number, and M denotes the max sentence length. When calculating the probability of each token, except for the context from source side x, the proposed SIM method employs historical and future information of intra- and inter-languages.

However, the parallel data is hard to collect. In this paper, we construct the trilingual training corpus by data augmenting. We first train 4 translation models (2 groups) M-1 and M-2, M'-1 and M'-2 on the bilingual training data (x_1, y_1) and (x_2, y_2). Then, these M-1 and M-2 translate input sentences x_2 and x_1 into pseudo training data (x_2, $y_1^{\triangle}$) and (x_1, $y_2^{\triangle}$). After that, we use model M'-1 and M'-2 to translate [$x_1 \cup x_2$], reverse, and get pseudo data [$\overleftarrow{y}_1^{\nabla} \cup \overleftarrow{y}_1^{\diamond}$] and [$\overleftarrow{y}_2^{\nabla} \cup \overleftarrow{y}_2^{\diamond}$], respectively. Therefore, we obtain the mixed parallel training data D =[(x_1, y_1, $\overleftarrow{y}_1^{\nabla}$, $y_2^{\triangle}$, $\overleftarrow{y}_2^{\nabla}$)]$\cup$[($x_2$, $y_1^{\triangle}$, $\overleftarrow{y}_1^{\diamond}$, y_2, $\overleftarrow{y}_2^{\diamond}$)] to train our model.

Experiments

Data & Settings

We evaluate our SimNMT method on two translation tasks, including English to German/French (briefly, En→De/Fr) and English to Chinese/Japanese (briefly, En→Zh/Ja) on IWSLT datasets[1]. The IWSLT.TED.tst2013 and IWSLT.TED.tst2015 are adopted as development set and test set, respectively.

En→De/Fr: The training sets contain 209.5K En→De and 236.7K En→Fr sentence pairs. We use the Moses (Koehn et al. 2007) tokenizer scripts[2] to tokenize English, German and French sentences[3]. Then, BPE method (Sennrich, Haddow, and Birch 2016) is adopted to encode the source sentences and the combination of all target sentences, respectively. The vocabulary sizes of both sides are limited to the most frequent 30.7K tokens.

En→Zh/Ja: The training sets contain 235.1K En→Zh and 226.8K En→Ja sentence pairs. We also use the scripts from Moses (Koehn et al. 2007) to tokenize English sentences, and urheen[4] to segment and tokenize Chinese and Japanese. After that, BPE (Sennrich, Haddow, and Birch 2016) is also used to encode the source sentences and the combination of all target sentences. The vocabulary sizes of both sides are limited to the most frequent 30.7K tokens.

We implement and evaluate the SimNMT with TensorFlow by modifying the tensor2tensor toolkit[5]. Specifically, we use 6 encoder and decoder layers Transformer with hidden size d_{model} = 512, 8 attention heads, 2,048 feed-forward inner layer size and $P_{dropout}$ = 0.1. The optimizer employs Adam method with parameters β_1 = 0.9, β_2 = 0.998 and ε = 10^{-9}. We adopt the same warm-up and decay settings as Vaswani et al. (2017). For testing, we use beam search with beam size k = 8 and allocate 2 beams for each type of translation hypotheses (specifically for 2 target languages in our experiments) with length penalty α = 0.6. The training and testing of all translation tasks are performed on single NVIDIA GTX 2080Ti GPU.

Baselines & Results

We compare the proposed model against the following advanced NMT and MNMT systems:

Standard Transformer (Vaswani et al. 2017): The classical NMT model, which employ the attention mechanism and predict target sentence from-left-to-right with standard beam search algorithm. Each model is trained with bilingual parallel corpus and can only translate one language pair.

[1] https://wit3.fbk.eu/

[2] https://github.com/moses-smt/mosesdecoder

[3] Since the original scripts can not split up the period from previous token if there are more than one sentence in a line, we modify the scripts to split them.

[4] http://www.nlpr.ia.ac.cn/cip/software.htm

[5] https://github.com/tensorflow/tensor2tensor

Method	**En-De/Fr**		**En-Zh/Ja**		**AVE**	**Δ**
	En-De	**En-Fr**	**En-Zh**	**En-Ja**		
Transformer (Vaswani et al. 2017)	26.48	38.61	16.24	15.29	24.16	-
Transformer (+pseudo) (Hoang et al. 2018)	27.14	39.59	17.36	16.21	25.08	+0.92
GNMT-Multi (Johnson et al. 2017)	28.29	40.30	17.86	17.06	25.88	+1.72
SB-NMT (Zhou, Zhang, and Zong 2019)	27.30	39.35	18.28	17.13	25.52	+1.36
SyncTrans (Wang et al. 2019)	26.58	39.95	17.80	17.69	25.51	+1.35
SimNMT	**28.63**	**41.41**	**18.68**	**17.18**	**26.48**	**+2.32**

Table 2: Translation performance on IWSLT datasets.

Transformer (+pseudo): The Transformer (Vaswani et al. 2017) adds pseudo corpus for training (Hoang et al. 2018). Specifically, for fair comparison, we group the experiments to En→De/Fr and En→Zh/Ja. In the En→De/Fr experiment, we first train the translation models M-(En→De) and M-(En→Fr) with parallel corpora En(De)-De and En(Fr)-Fr, respectively. Then, translate En(Fr) and En(De) with these two models, and obtain the pseudo parallel corpora En(Fr)-De(pseudo) and En(De)-Fr(pseudo). Finally, we combine the original corpora En(De)-De and En(Fr)-Fr with the pseudo corpora En(Fr)-De(pseudo) and En(De)-Fr(pseudo), respectively, and then train the Transformer as Transformer (+pseudo). The same goes for En→Zh/Ja.

GNMT-Multi (Johnson et al. 2017): The Google's multilingual NMT system, which also employs standard beam search algorithm but enables multilingual translations by adding direction tags to source language sentences. All training data of different language pairs are mixed together.

SB-NMT (Zhou, Zhang, and Zong 2019): A synchronous bidirectional NMT model, which decodes a target sentence from both left side and right side in a synchronous way. For each source sentence, it needs an extra target sentence in reverse order with the same semantics for training.

SyncTrans (Wang et al. 2019): This work simultaneously translate one source language into two target languages with one encoder and two different decoders. It requires triple parallel corpus for training.

The training corpora are prepared according to the schemes in respective papers. The translation performance is evaluated by BLEU (Papineni et al. 2002), and all experiments are case-insensitive. For Transformer, Transformer (+pseudo) and SB-NMT, each model translates one language pair; while for GNMT-Multi, SyncTrans and proposed SimNMT, 2 groups of multilingual translation experiments are carried out, En→De/Fr and En→Zh/Ja, respectively.

Table 2 shows the overall translation results of En→De/Fr/Zh/Ja on the IWSLT datasets. For the average (AVE) BLEU of 4 target languages, SimNMT obviously outperforms existing methods. It achieves +2.32 average BLEU score higher than standard Transformer, and +0.60 BLEU improvement over GNMT-Multi. SimNMT significantly outperforms other methods on all En-De, En-Fr and En-Zh individual translation tasks, and obtains comparable results with the SyncTrans on En-Ja translation task.

Fusion Para		λ_1	λ_2	λ_1	λ_2	λ_1	λ_2
		1.0	0.1	0.7	0.1	0.5	0.1
Linear		32.68		34.08		33.34	
Non-linear	**ReLU**	34.28		35.02		34.26	
	tanh	34.01		34.79		34.13	
Gate		33.24					

Table 3: Experiment results with different fusion parameters.

Analysis

We conduct analyses on several aspects to better understand our model.

Fusion Scheme: We compared different attention fusion functions (FF) and parameters. The fusion functions tested in this paper are as follows:

$$\begin{aligned}\vec{\mathrm{H}} &= \mathrm{FF}(\vec{\mathrm{H}}_{intra}^{history}, \vec{\mathrm{H}}_{intra}^{future}, \vec{\mathrm{H}}_{inter}^{history}, \vec{\mathrm{H}}_{inter}^{future}) \\ &= \lambda_1 \cdot \vec{\mathrm{H}}_{intra}^{history} + \lambda_2 \cdot AF(\vec{\mathrm{H}}_{intra}^{future}) + \\ &\lambda_1 \cdot AF(\vec{\mathrm{H}}_{inter}^{history}) + \lambda_2 \cdot AF(\vec{\mathrm{H}}_{inter}^{future})\end{aligned} \tag{10}$$

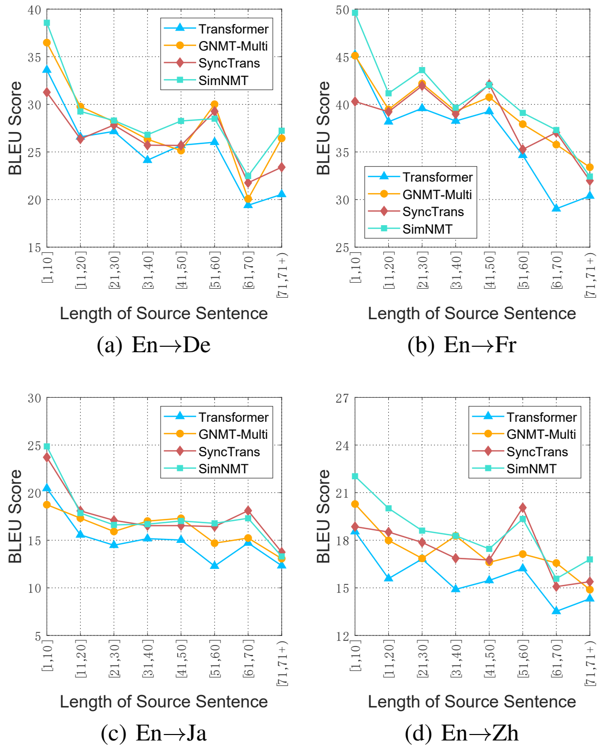
where the activation functions (AF) include linear interpolation (Linear), non-linear interpolation (Non-linear) of rectified linear unit (ReLU) and tanh function, and gate mechanism (Gate). Many empirical experiments of En→De/Fr show that different fusion parameters have a great impact on translation performance (Table 3 shows the average BLEU scores of De and Fr). When the parameters of the ReLU function λ_1=0.7, λ_2=0.1, we get the best result. The gate mechanism shows worse performance than most fixed parameter methods. Since the fusion scheme is very important to the result, the better fusion mechanism is a direction worth studying in the future.

Effect of Sentence Length: We calculated the average translation quality of different sentence lengths (by grouping similar lengths together) to observe the performance of different methods. Figure 6 shows SimNMT reaches better BLEU scores in general (sentence length in the range of [1,70]) compared with other methods.

Translation Balance: Translation that only relies on historical information (L2R) faces the translation unbalance issue, i.e., the translation quality of the second half of target sentence reducing. By introducing future information (R2L), the model is expected to get more accurate translation of the suffix in a sentence. As shown in Figure 7, we count the number of translated sentences whose first and last 4 tokens are exactly the same with the references. Compared with the

Source	I have to create a framework that you then fill in with your imagination .	Now , you may have noticed there are some people missing there : the rest of the team .
Reference	来 建立 起 一 个 框架 然后 你运用 想像力 去 充实 它	现在 你 可能 注意 到 球场 上 少 了 些 什么 人 少 了 球队 的 其他 成员
	To build a framework and then you use your imagination to enrich it	Now you may notice what is missing on the court. Other members of the team are missing.
Transformer	我 需要 创造 一 个 框架 ， 一 个 充满 想象力 的 框架 。	也许 你 已经 注意 到 有 一些 人 失踪 了 ， 其他 的 团队 。
	I need to create a framework, a framework full of imagination.	Maybe you have noticed that some people are missing, other teams.
SyncTrans	我 得 创造 一 个 你 可以 用 你 的 想象力 填满 的 框架 。	现在 ， 你们 可能 已经 注意 到 有些 人 在 这里 失踪 了 ： 其他 团队 。
	I have to create a frame that you can fill with your imagination.	Now, you may have noticed that some people are missing here: other teams.
SimNMT	我 得 创造 一 个 构架 ， 让 你 可以 用 想象力 来 填充 。	现在 ， 你们 可能 已经 注意 到 有些 人 失踪 了 ： 团队 的 其他 人 。
	I have to create a framework that you can fill with imagination.	Now, you may have noticed that some people are missing: the rest of the team.

Table 4: The En→Zh translation examples of Transformer, SyncTrans and proposed SimNMT. The Chinese sentences are translation results by different methods, while the below English sentences are the back translations.



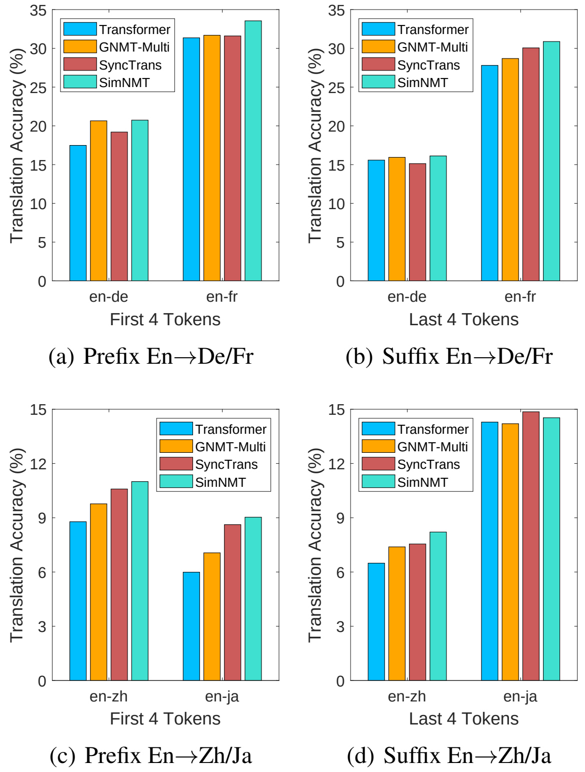
(a) En→De

(b) En→Fr

(c) En→Ja

(d) En→Zh

Figure 6: Translation qualities (BLEU score) of different lengths of the source sentences.



(a) Prefix En→De/Fr

(b) Suffix En→De/Fr

(c) Prefix En→Zh/Ja

(d) Suffix En→Zh/Ja

Figure 7: Translation accuracies of the prefixes and suffixes.

methods without future information (Transformer, GNMT-Multi and SyncTrans), translation accuracies of both sides of SimNMT are generally improved.

Case Study: Table 4 gives two case study examples to better understand how the proposed SimNMT model outperforms than other models. In the first example, SimNMT gives more appropriate translation. In the second example, SimNMT provides the only correct translation suffix, indicating that the proposed method improves the translation performance, such as translation suffixes.

Conclusions

In this paper, we propose a synchronous interactive multilingual NMT (SimNMT) model that translates one source language into different target languages simultaneously and interactively. The synchronous cross-interactive decoder, which can take full advantage of four types of information (historical and future information of intra-language and inter-languages), predicts output for each target language with proposed synchronous interactive multilingual inference algorithm. To our best knowledge, this is the first attempt to integrate above four types of information into a single NMT model. The experiments demonstrate that the proposed approach obtains significant improvements over competitive bilingual NMT models and multilingual NMT models. In future work, we plan to extend our method on more than two target languages, and explore better ways to fuse and utilize language characteristics to further improve multilingual translation performance.

Acknowledgments

The research work descried in this paper has been supported by the Natural Science Foundation of China under Grant No. U1836221 and Beijing Academy of Artificial Intelligence BAAI2019QN0504.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Dong, D.; Wu, H.; He, W.; Yu, D.; and Wang, H. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, volume 1, 1723–1732.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 1243–1252.

Hoang, V. C. D.; Koehn, P.; Haffari, G.; and Cohn, T. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation (WNMT 2018)*, 18–24.

Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viegas, F.; Wattenberg, M.; Corrado, G.; et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics (TACL)* 5: 339–351. ISSN 2307-387X.

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; and Zens, R. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 177–180.

Liu, L.; Utiyama, M.; Finch, A.; and Sumita, E. 2016. A-greement on target-bidirectional neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 411–416.

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

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1715–1725.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 3104–3112.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 5998–6008.

Wang, Y.; Zhang, J.; Zhai, F.; Xu, J.; and Zong, C. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2955–2960.

Wang, Y.; Zhang, J.; Zhou, L.; Liu, Y.; and Zong, C. 2019. Synchronously generating two languages with interactive decoding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3341–3346.

Zhang, J.; Zhou, L.; Zhao, Y.; and Zong, C. 2020. Synchronous bidirectional inference for neural sequence generation. *Artificial Intelligence* 281: 103234. ISSN 0004-3702.

Zhang, X.; Su, J.; Qin, Y.; Liu, Y.; Ji, R.; and Wang, H. 2018. Asynchronous bidirectional decoding for neural machine translation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 5698–5705.

Zhou, L.; Zhang, J.; and Zong, C. 2019. Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics (TACL)* 7: 91–105. ISSN 2307-387X.

Zhou, L.; Zhang, J.; Zong, C.; and Yu, H. 2019. Sequence generation: from both sides to the middle. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 5471–5477.