Synchronous Inference for Multilingual Neural Machine Translation

Qian Wang, Jiajun Zhang, *Senior Member, IEEE*, and Chengqing Zong, *Senior Member, IEEE*

***Abstract*—Multilingual neural machine translation allows a single model to translate between multiple language pairs, which greatly reduces the cost of model training and receives much attention recently. Previous studies mainly focus on training stage optimization and improve positive knowledge transfer among languages with different levels of parameter sharing, but ignore the multilingual knowledge transfer during inference although the translation in one language may help the generation of other languages. This work enhances knowledge sharing among multiple target languages in the inference phase. To achieve this, we propose a synchronous inference method that can simultaneously generate translations in multiple languages. During generation, the model that predicts the next word of each language not only based on source sentence and previously predicted segments, but also based on predicted words of other target languages. To maximize the inference stage knowledge sharing, we design a cross-lingual attention module which allows the model to dynamically selects the most relevant information from multiple target languages. The synchronous inference model requires multi-way parallel training data which is scarce. We therefore propose to adopt multi-task learning to incorporate large-scale bilingual data. We evaluate our method on three multilingual translation datasets and prove that the proposed method significantly improve the translation quality and the decoding efficiency compared to strong bilingual and multilingual baselines.**

***Index Terms*—Multilingualism, neural machine translation, synchronous inference.**

I. Introduction

NEURAL machine translation has achieved great success during the past few years [1]–[5]. In neural machine translation, the underlying encoder-decoder based neural network directly models the mapping from a source language to a target language, which provides simplicity of implementation and better translation quality compared to statistic methods [6]–[8]. In parallel to the development of neural machine translation, the end-to-end training paradigm also makes it feasible to build a multilingual neural machine translation (MNMT) system by simply extending the source and/or the target side languages [9]–[12]. Multilingual neural machine translation handles multiple translation directions within a shared neural model, which significantly reduces the cost of offline training and online deploying, and improves the translation accuracy due to the positive knowledge transfer with the shared parameters [13]–[15].

The dominant paradigm for building a multilingual neural machine translation system uses a single shared model for all translation directions and the universal model is jointly trained on a concatenated dataset that contains parallel sentences of various language pairs [10], [11]. Previous studies mainly focus on exploring the optimal parameter sharing strategy [16]–[18] that can maximize the positive knowledge transfer among languages by utilizing the complementary information of different languages in the *training* process. Recently, [19], [20] attempt to extend the paradigm by enabling *inference* stage knowledge transfer. They build a multilingual translation model that can translate a source language into two different target languages simultaneously and interactively, which utilizes the information from the different languages in the *inference* process.

However, the existing synchronous generation methods still face several practical issues [19], [20]. First, previous studies only consider generating two languages simultaneously and lacks generalization ability to multiple target language translation, although multiple target languages can provide more effective guidance for generating a specific target language. As shown in Fig. 1, when predicting the Chinese word “ the generated words “ “linguagem” in Portuguese and “ dn BAqA” in Hindi can provide more contextual information. Incorporating multiple target languages also brings robustness, e.g., when the word “linguagem” is under-translated, the other two languages can still help the prediction of the Chinese word “ Second, previous synchronous generation methods use a predefined weight to manually balance the contributions of the two target languages. When multiple target languages are involved, the predefined weight makes it unfeasible to dynamically select the most relevant information for each language, since different languages may have different effects on the translation of a specific language. Third, the training process of previous synchronous generation model only depends on limited multi-way parallel data and can not exploit the existing large-scale bilingual parallel training data.

In this work, we propose a fully synchronous inference method for multilingual neural machine translation that can generate multiple target sentences in different languages simultaneously and interactively. In the generation process, the

Manuscript received January 9, 2022; revised April 23, 2022; accepted May 19, 2022. Date of publication May 27, 2022; date of current version June 6, 2022. This work is supported by the Natural Science Foundation of China under Grants 62122088, U1836221, and 62006224. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sakriani Sakti. *(Corresponding author: Jiajun Zhang.)*

Qian Wang, Jiajun Zhang, and Chengqing Zong are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: qian.wang@nlpr.ia.ac.cn; jjzhang@nlpr.ia.ac.cn; cqzong@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TASLP.2022.3178241

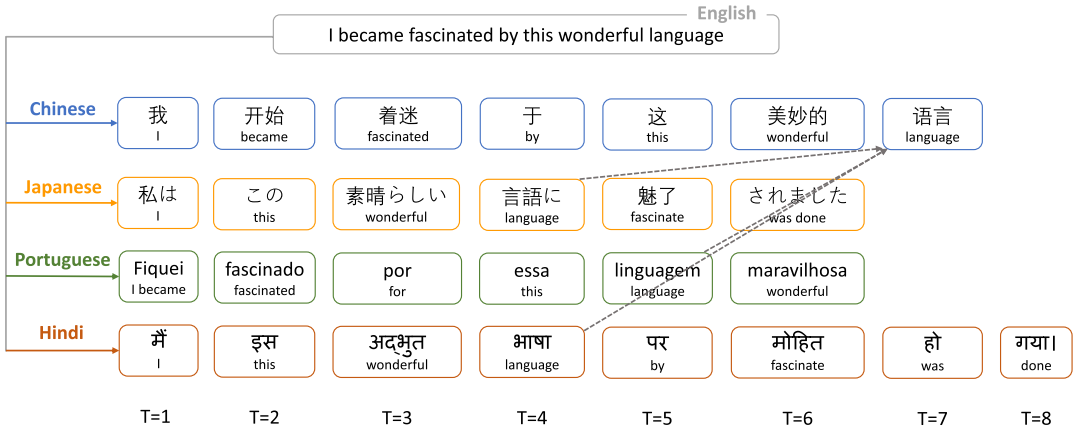

Fig. 1. An example of an English sentence translated into 4 target languages, including Chinese, Japanese, Portuguese, and Hindi. The 4 target sentences interact with each other during generation to facilitate the word prediction.

model that predicts the next word relies on not only the source and previously generated words, but also the predicted tokens from other languages. To utilize complementary knowledge of different target languages during generation, we design a **cross-lingual attention** module that can dynamically select the most relevant part from target sentences of multiple auxiliary languages to guide the generation of language of interest. In this way, our method can generate translation in multiple languages simultaneously and allows mutual boosting among those target languages. Furthermore, in addition to the limited multi-way parallel training data used in previous synchronous generation methods, we apply multi-task learning to incorporate the existing large-scale bilingual parallel dataset. The two tasks, including vanilla multilingual translation and synchronous generation, are jointly trained on the multi-way parallel data and the bilingual parallel data.

We verify the effectiveness of our method through extensive experiments on three English-centered multilingual datasets, including the small-scale multi-way aligned IWSLT'14 dataset, the large-scale multi-way aligned UN dataset, and the WMT'17 dataset that includes both multi-way aligned data and bilingual parallel data.

Our main contributions are summarized as follows:

1) We propose a synchronous inference method for multilingual NMT that can translate into multiple languages simultaneously.
2) We propose cross-lingual attention that allows the generation process to dynamically select relevant information from multiple target languages.
3) In addition to the limited multi-way parallel training data, we also propose to adopt multi-task learning to exploit large-scale bilingual data for training our model.

II. Related Work

A. Multilingual Neural Machine Translation

Multilingual neural machine translation aims to handle translation between multiple language pairs within a single model [21]. Compared to their bilingual counterparts, multilingual models greatly reduce the parameter scale while preserving similar translation quality. In recent years, researchers have proposed various model architectures with different levels of parameter sharing for building a multilingual neural machine translation model. [9] applies multi-task learning for training a one-to-many multilingual model with a completely shared encoder and separate decoders. [22] uses a shared decoder for multi-source translation that generates translation based on multiple semantically identical sentences from different languages. Beside shared encoder or shared decoder, shared attention with separate encoders and separate decoders has also been proposed to enable many-to-many multilingual translation [16]. These methods share modules for the same language of different tasks and reduce the parameter scale from $\mathcal{O}(m \times n)$ to $\mathcal{O}(m + n)$, where m and n are the number of source languages and target languages respectively.

The success in reducing parameter scale with shared modules motivates a more aggressive sharing strategy that further reduces the parameter scale to $\mathcal{O}(1)$. [11] proposes to use a single completely shared model borrowed from bilingual translation for all translation directions. They concatenate bilingual corpora of different language pairs and append a special language token to each input sentence to indicate the target language. In the meantime, [10] proposes a similar model but maintains separate vocabularies for each language. The shared many-to-many multilingual model brings positive knowledge transfer and better translation quality for low-resource languages, and thus becomes the standard paradigm in multilingual neural machine translation.

B. Synchronous Inference

Most neural machine translation models follow the autoregressive generation style which predicts the target sentence from left to right and one token at a time. Beam search decoding strategy is widely used to find a translation that approximately maximizes the conditional log probability. However, the fixed left-to-right decoding order can only utilize past information for predicting the current word and researchers find that the

future information may also provide valuable guidance since the translations generated by left-to-right decoding order and the right-to-left decoding order are complementary to each other [23], [24].

To effectively exploit the future information during decoding, [25], [26] propose a synchronous bidirectional decoding method for sequence generation tasks like neural machine translation. In synchronous generation, the model produces two translations following left-to-right order and right-to-left order respectively. The two translations are generated simultaneously and interactively with a special synchronous attention module and enhanced beam search algorithm. [27] further improves the decoding efficiency by generating only half translations for both left-to-right order and right-to-left order and concatenating the two halves to obtain the final translation.

The synchronous generation methods prove the effectiveness in acquiring knowledge from a different auxiliary task during inference. Researchers then investigate the possibility of incorporating different tasks [28], such as summarization [26], speech recognition and speech-to-text translation [29]. [19], [20] adopt the synchronous generation to multilingual translation and allow the model to generate translations in two languages simultaneously and interactively.

However, the existing synchronous interactive generation methods can only generate two targets at the same time and lack generalization ability to multiple target languages, although multiple target languages can provide more effective knowledge for translation and improve the robustness for synchronous inference. From this viewpoint, we propose a synchronous inference method for multilingual neural machine translation to generate translations in multiple languages simultaneously.

III. Background

In this work, we build our model based on the Transformer architecture [5]. Equipped with beam search decoding algorithm, the Transformer model achieves state-of-the-art results for machine translation and becomes the de facto standard for building a neural machine translation system in the research community and industry. In this section, we will briefly introduce the Transformer model and the beam search decoding algorithm.

1) The Transformer Model: The transformer model [5] follows the traditional encoder-decoder framework that consists of an encoder and a decoder. Given an input sentence $\mathbf{x} = (x_1, \ldots, x_n)$ that contains a list of tokens, the encoder network first transforms x into a sequence of continuous representation $\mathbf{z} = (z_1, \ldots, z_n)$. The decoder then generates the translation $\mathbf{y} = (y_1, \ldots, y_m)$ based on the representation $\mathbf{z}$ one word at a time. As shown in Fig. 2, both the encoder and the decoder are stacked with N identical layers. In the encoder, each layer has two sub-layers: the multi-head self-attention sub-layer and the position-wise feed-forward network. Residual connection [30] and layer normalization [31] are employed for each of the two sub-layers. In the decoder, except for the multi-head self-attention and the feed-forward network in each encoder layer, the decoder also inserts an additional multi-head attention module

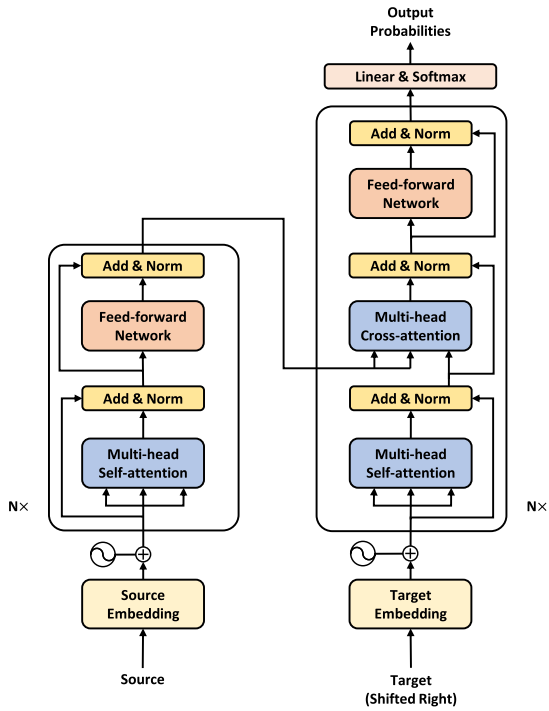

Fig. 2. The overview of the Transformer model.

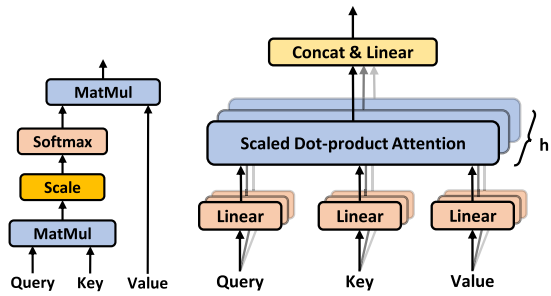

Fig. 3. The scaled dot-product attention (left) and the multi-head attention (right).

that attends to the continuous representation z generated by the encoder.

2) Multi-head Attention: The power of the Transformer model relies on the multi-head attention that jointly attend to different positions from different representation sub-spaces. As shown in Fig. 3, the multi-head attention module takes a set of queries, keys, and values as input. For the self-attention sub-layer in both the encoder and the decoder, the queries, the keys, and the values are identical and come from the output hidden states of the previous layer. For the cross-attention sub-layer in the decoder, the queries are from previous layers while the keys and the values are from the continuous representation $\mathbf{z}$ from the encoder. Formally, the multi-head attention calculates the attention from different representation sub-spaces through different attention heads and concatenates the representations of each head to obtain the output:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{1}$$

where W_i^Q, W_i^K, W_i^V and W^O are parameter matrices of linear projections.

The attention used in the Transformer is called Scaled Dot-Product Attention (Fig. 3), in which each head is computed as a weighted sum of the values, and the weights are obtained by a softmax function that operates on the scaled dot-product of the queries and the keys:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

3) Beam Search Decoding: In the inference stage, the trained neural machine translation model predicts the target sentence from left to right and one token at a time. Beam search decoding algorithm is used to find a translation that approximately maximizes the conditional log-probability $\sum_t \log p(y_t|\mathbf{x}, y_{<t})$, where $y_{<t}$ is the previously generated tokens. In beam search with size B, only B hypotheses are stored and extended to avoid full space searching. At decoding step t, the algorithm first selects the top-B words with highest probabilities for each of the B hypotheses with length $t-1$, which results in $B \times B$ candidates. Then, the top-B candidates with length t are selected and stored. The decoding process continues until a special end-of-sentence token appears which indicates that the translation of the current source sentence is finished.

IV. Our Approach

In this section, we will introduce the proposed synchronous inference framework for multilingual machine translation. We first introduce the model that can generate translations in multiple languages simultaneously (§IV-A). The model consists of a specifically designed attention called cross-lingual attention that enables information interaction of multiple target languages in the decoder (§IV-B). To fit in the synchronous generation, we extend the beam search algorithm that can generate tokens in multiple languages simultaneously (§IV-C). Finally, we introduce the training strategy for the model and the multi-task training schema to utilize both limited multi-way parallel data and large-scale bilingual data (§IV-D).

A. The Synchronous Inference Model

The goal of our model is to translate a given source sentence $\mathbf{x} = (x_1, \ldots, x_n)$ into multiple target languages $\{\mathbf{y}^1, \ldots, \mathbf{y}^L\}$ simultaneously, where $\mathbf{y}^l = (y_1^l, \ldots, y_m^l)$ is the translation of the l-th target language and L is the total number of target languages. We build our model with the Transformer architecture [5] and the overview of the proposed model is illustrated in Fig. 4.

1) The Encoder: The encoder in our model is identical to the Transformer encoder that maps the source sentence $\mathbf{x}$ into a continuous representation $\mathbf{z}$. In previous multilingual neural machine translation methods, the encoder appends a special language ID to each input sentence to indicate the target language: $\mathbf{x} = (\langle\text{lang}\rangle, x_1, \ldots, x_n)$. However, the language token in the source sentence makes the representation $\mathbf{z}$ target-dependent, i.e., the representation is correlated with one specific target language. In our model, since multiple target languages are generated simultaneously, we remove the language tokens in the source sentence to make the source representation target-agnostic.

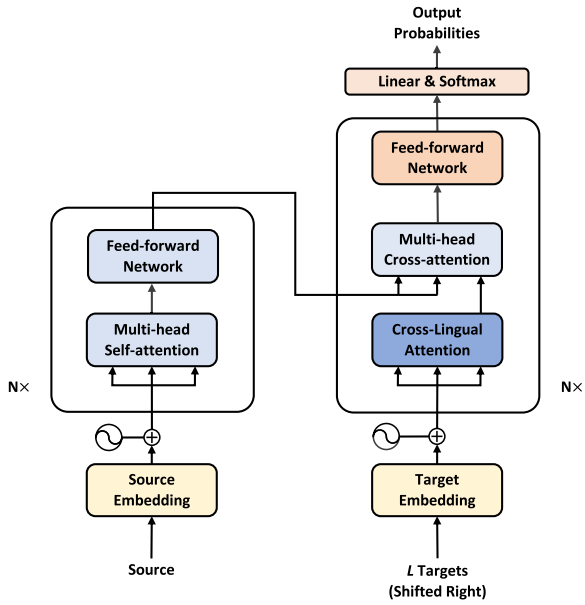

Fig. 4. The overview of the proposed synchronous generation model. We omit the details of residual connection and layer normalization for simplicity.

2) The Decoder: Given the encoder representation $\mathbf{z}$ and the previously generated tokens $y_{<t}^l$ for each target language, the decoder in our model is responsible for predicting the next target token y_t^l. As shown in the right part of Fig. 4, the input of our decoder consists of the already generated tokens of L target languages:

$$y = \begin{bmatrix} y_1^1 & \cdots & y_{t-1}^1 \\ \vdots & \ddots & \vdots \\ y_1^L & \cdots & y_{t-1}^L \end{bmatrix} \tag{3}$$

where t is the current decoding step. The decoder consists of N identical layers and each layer includes three sub-layers: the feed-forward network and the multi-head cross-attention are borrowed from the Transformer decoder, while the cross-lingual attention is designed for synchronous generation and cross-lingual interaction. The inner representation of the decoder $h \in R^{L \times m \times d}$ consists of hidden states of all the L target languages, where m is the sentence length and d is the hidden dimension. From the final representation produced by the last decoder layer, the model predicts the next words for each of the L languages simultaneously.

B. The Cross-Lingual Attention

We replace the multi-head self-attention in the Transformer decoder with a specifically designed cross-lingual attention

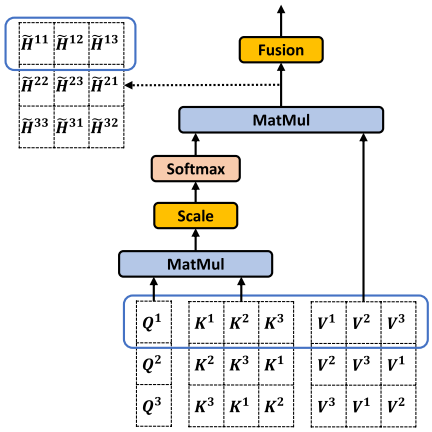

Fig. 5. The attention calculation of $L = 3$ target languages with the proposed cross-lingual attention based on scaled dot-product attention.

module to enable interaction among different target languages. This module takes a set of queries Q, keys K, values $V \in R^{L\times m\times d}$ with size as input, and generates hidden representations for each language simultaneously.

Different from the scaled dot-product attention in the Transformer decoder, we allow each target language to not only attend to itself but also attend to other target languages during the calculation of attention. To achieve this, we first obtain the attention between each two languages. As shown in Fig. 5, we reorder the input to facilitate parallel computation:

$$\mathcal{Q} = \left[Q^1 \dots Q^L\right]^T \tag{4}$$

$$\mathcal{K} = \begin{bmatrix} K^1 & \cdots & K^L \\ \vdots & K^{[(i+j-2) \bmod L]+1} & \vdots \\ K^L & \cdots & K^{L-1} \end{bmatrix} \tag{5}$$

$$\mathcal{V} = \begin{bmatrix} V^1 & \cdots & V^L \\ \vdots & V^{[(i+j-2) \bmod L]+1} & \vdots \\ V^L & \cdots & V^{L-1} \end{bmatrix} \tag{6}$$

where Q^i, K^i and V^i are queries, keys and values of the i-th language from a specific sub-space.

The attention of the i-th language to the j-th language $\tilde{H}^{ij}$ can be calculated as:

$$\begin{aligned} \tilde{H}^{ij} &= \text{Attention}\left(Q^i, K^j, V^j\right) \\ &= \text{Softmax}\left(\frac{Q^i (K^j)^T}{\sqrt{d_k}}\right) V^j \end{aligned} \tag{7}$$

The pair-wise attention are then combined with a fusion function to generate the final representation for each language:

$$H^i = \text{Fusion}\left(\tilde{H}^{i1}, \tilde{H}^{i2}, \dots, \tilde{H}^{iL}\right) \tag{8}$$

For the fusion functions, we first adopt the linear interpolation and nonlinear interpolation as in other synchronous generation methods [25], and extend the two functions to fit in multiple target languages. We then propose attention-based fusion function that dynamically selects the most relevant information from different languages.

Algorithm 1: Synchronous Beam Search Algorithm.

Input: Encoder $f(\cdot)$, decoder $g(\cdot)$, input sequence $\mathbf{x} = (x_1, \dots, x_n)$, beam size B, maximum length of output sequence M, target language IDs ids, vocabulary size V, number of target languages L.
Output: Translations in multiple languages $\mathbf{y}$.

```
z ← f(x)
Initialize hypotheses y with size L × B × M and fill y with ⟨pad⟩.
Initialize scores (log-probability) s with size L × B × M and fill s with 0.
for l = 0 to L do
    y[l, 0, 0] ← ids[l]        ▷ Initialize the first token of the first beam as language token.
end for
for m = 1 to M do
    p = g(y[:, :, : m], z)     ▷ p ∈ R^{L×B×|V|} is the log-probability of each word at m-th step.
    p = p + s[:, :, m]         ▷ Accumulated the log-probabilities of m-th step and previous steps.
    for l = 0 to L do
        Select top-B candidates from p[l].
        Expand tokens y[l] with those candidates.
        Expand scores s[l] with those candidates.
    end for
end for
```

1) Linear Interpolation: Intuitively, the importance of self-attention $\tilde{H}^{ii}$ is more important than the attention to other languages $\{\tilde{H}^{ij} | i \neq j\}$ for generating the hidden representation of the i-th language. We use linear interpolation to control the importance of attention to different languages:

$$H^i = \tilde{H}^{ii} + \lambda \times \sum_{j\neq i} \tilde{H}^{ij} \tag{9}$$

where λ is a hyper-parameter to balance the importance from other languages for generating i-th language.

2) Nonlinear Interpolation: To better distinguish the representation of self-attention and the attention to other languages, we add a nonlinear activation function $AF(\cdot)$ to other languages:

$$H^i = \tilde{H}^{ii} + \lambda \times \sum_{j\neq i} AF\left(\tilde{H}^{ij}\right) \tag{10}$$

where AF can be any nonlinear activation functions such as tanh, ReLU or Sigmoid.

3) Attention-based Fusion: The previous two fusion methods treat the target languages equally. However, different languages may have different effects on the translation of a specific language. Therefore, we propose an attention-based fusion function that allows one language to dynamically select relevant information from all languages other than using fixed weight λ in linear or nonlinear interpolation. The attention-based fusion

is calculated as:

$$H^i = \text{Attention}\left(\begin{bmatrix}\tilde{H}^{i1}\\ \vdots \\ \tilde{H}^{iL}\end{bmatrix} W^Q, \begin{bmatrix}\tilde{H}^{i1}\\ \vdots \\ \tilde{H}^{iL}\end{bmatrix} W^K, \begin{bmatrix}\tilde{H}^{i1}\\ \vdots \\ \tilde{H}^{iL}\end{bmatrix} W^V\right),$$
$$\text{where Attention}\,(H_Q, H_K, H_V) = \text{Softmax}\left(\frac{H_Q (H_K)^T}{\sqrt{d_k}}\right) H_V \tag{11}$$

After fusion, the representation of each language contains information not only from itself, but also from other languages. The output of the cross-lingual attention module $H = [H^1, \ldots, H^L]$ is the combination of representations of all languages.

C. Synchronous Beam Search

Given a trained translation model and a source sentence $\mathbf{x}$, the beam search algorithm is used to generate a target sequence $\mathbf{y}$. We extend the traditional beam search to generate multiple target sequences in different languages simultaneously.

Our synchronous beam search algorithm is shown in Algorithm 1. After obtaining the representation $\mathbf{z}$ of the input source sentence $\mathbf{x}$ (Line 1), we initialize the hypotheses token cache and the score cache (Line 2-3). Different from the bilingual translation model that use a special $\langle\text{BOS}\rangle$ to initialize the first token of the hypothesis, we initialize the first token of each hypothesis with a language ID corresponding to the target language (Line 4-6). At each decoding step, the decoder predicts the output distributions of each language based on the encoder output and the previously generated tokens of all target languages (Line 8). The algorithm then selects the best top-B words for each language and updates the cache for each language (Line 10-14). The generation process continues until the translation of all languages are finished or the maximum sequence length is reached. By using the synchronous beam search with our model, the decoding process of each language can interact with each other in one beam search process.

D. Model Training

Since our synchronous generation model translates a source sentence into multiple target languages in parallel, the model training requires multi-way aligned parallel data, i.e., semantically identical sentences are given in all languages. Given the training data $\mathcal{D}_{\text{multi-way}} = \mathcal{D}^{S\to\{T_1,\ldots,T_L\}}$, the training objective is:

$$\mathcal{L}_1(\theta) = \underbrace{\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_{\text{multi-way}}} \sum_{i=1}^{M}\sum_{l=1}^{L} -\log p\left(\mathbf{y}_i^l | \mathbf{x}, \mathbf{y}_{<i}^1, \ldots, \mathbf{y}_{<i}^L\right)}_{\text{Multi-way Loss}} \tag{12}$$

where L is the number of target languages and M is the sentence length. Since different target sentences may have different lengths, we pad the shorter target sentences with a special $\langle\text{PAD}\rangle$ token to match the maximum sentence length $M = \max(\text{len}(\mathbf{y}^1), \ldots, \text{len}(\mathbf{y}^L))$, and the loss of the $\langle\text{PAD}\rangle$ token will be ignored during training.

However, the training requires multi-way aligned parallel data, which is much scarcer than bilingual parallel data. To exploit the existing large-scale bilingual parallel data, we adopt multi-task learning [32] to facilitate the training of our synchronous generation model. We design two tasks that are jointly optimized: the synchronous generation task and the multilingual translation task. The synchronous generation task aims to translate one source sentence into multiple target languages as described in (12), while the multilingual translation task aims to translate one source sentence into one target language at a time.

Suppose we have a bilingual parallel dataset $\mathcal{D}_{bi} = \{\mathcal{D}_{bi}^{S\to T_1}, \ldots, \mathcal{D}_{bi}^{S\to T_L}\}$ for different language pairs $S \to T_l$, the training objective of the multilingual translation task is to minimize the negative log-likelihood on different bilingual dataset $\mathcal{D}_{bi}^{S\to T_l}$ in $\mathcal{D}_{bi}$:

$$\mathcal{L}_2(\theta) = \underbrace{\overbrace{\sum_{l=1}^{L} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_{bi}} \sum_{i=1}^{M} -\log p(\mathbf{y}_i | \mathbf{x}, \mathbf{y}_{<i})}^{\text{Bilingual Loss}}}_{\text{Multilingual Loss}} \tag{13}$$

The two tasks, including the vanilla multilingual translation and the synchronous generation, are jointly trained on the multi-way parallel data and the bilingual parallel data. We balance the importance of the two tasks with parameter α that controlled by a sigmoid decay function:

$$\alpha = \frac{1}{1 + \exp(-10\,k/T)}$$
$$\mathcal{L}(\theta) = (1-\alpha)\mathcal{L}_1(\theta) + \alpha\mathcal{L}_2(\theta) \tag{14}$$

where k is the current training step and T is the total training step. In the beginning, the model is trained with both the multilingual translation objective and the synchronous objective to utilize the large-scale bilingual parallel dataset. As the training proceeds, the weight of the synchronous generation task gradually increases to strengthen the interaction among different target languages. To stabilize convergence, we first pretrain a multilingual model on the entire dataset and continue training with the multi-task learning objective.

V. Experimental Setup

A. Dataset

We evaluate the proposed method on three public English-centric datasets: the IWSLT'14 dataset [33], the UN dataset [34], and the WMT'17 dataset [35].

The IWSLT'14 dataset consists of bilingual parallel data of English to 6 target languages: English $\to$ {Spanish, Dutch, Portuguese, Romanian, Russian, Chinese} (briefly EN $\to$ ES/NL/PT/RO/RU/ZH}). We use a combination of *dev2010*, *tst2010*, *tst2011* and *tst2012* as validation set and *tst2014* as test set. For the training data and the validation data, we extract 7-way parallel data for EN/ES/NL/PT/RO/RU/ZH [22] by comparing the source (English) side of each bilingual dataset and selecting sentences pairs if their source side sentences are identical. For the test set, we use the original data without multi-way parallel selection. Table I summarizes the data statistics for

TABLE I
STATISTICS OF THE IWSLT'14 DATASET USED IN OUR EXPERIMENTS

	EN→ES	EN→NL	EN→PT	EN→RO	EN→RU	EN→ZH	7-way
Train	180,850	167,943	171,903	182,141	178,165	179,901	150,970
Valid	5,593	5,389	5,388	5,585	5,537	5,099	4,742
Test	2,504	2,457	2,351	2,463	2,331	2,329	-

TABLE II
STATISTICS OF THE WMT'17 DATASET USED IN OUR EXPERIMENTS

	EN→DE	EN→LV	EN→FI	4-way
Train	5,852,458	4,461,720	2,634,433	647,152
Valid	5,593	5,389	5,388	5,585
Test	2,504	2,457	2,351	2,463

training, validation and testing. We segment the Chinese data with Jieba tokenizer[1] and segment the other languages with Moses tokenizer[2] [36]. We use byte-pair encoding (BPE) [37] to encode all the training data and learn a shared subword vocabulary with a size of 32 k.

The UN dataset is composed of United Nations documents in six official languages of the United Nations: Arabic, Chinese, English, French, Russian, and Spanish (briefly AR, ZH, EN, FR, RU, ES). The dataset consists of 11 M 6-way parallel data for training, 4,000 6-way parallel data for validation and 4,000 6-way parallel data for testing. The Chinese data is segmented with Jieba tokenizer while the other 5 languages are segmented with Moses tokenizer. BPE is also used to obtain the shared vocabulary of about 64 K subword tokens.

For the WMT'17 dataset, we use the English $\rightarrow$ {German, Latvian, Finnish} (briefly EN $\rightarrow$ DE/LV/FI) bilingual datasets. We extract the 4-way parallel data for English, Latvian, German and Finnish and train the synchronous generation model. We create the 4-way validation data by combining the English side of *newsdev2017* and *newstest2016*, and translate the English sentences into German, Latvian, and Finnish with a pretrained bilingual translation model. The *newstest2017* is used for testing. As shown in Table II, the 4-way parallel data size is quite smaller than the original bilingual data. To utilize the entire dataset, we also incorporate the original bilingual data with multi-task learning described in Section IV-D. Similar to the other two datasets, we use Moses tokenizer and apply BPE with a shared vocabulary of 48 K tokens.

B. Training and Evaluation Details

We implement the proposed synchronous generation model using fairseq toolkit [38]. We use Adam optimizer [39] with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The learning rate is set to 7×10^{-4} and inverse square root decay strategy with a warmup of 8,000 steps is used. We adopt dropout [40] of $p = 0.1$ during training. All models are trained and evaluated on a single Nvidia GTX 3090 GPU. For the model sizes, we adopt two different settings for the three datasets in our experiment.

For the smaller IWSLT'14 dataset, we follow the *transformer_iwslt_de_en* setting,[3] which includes 6 encoder layers and 6 decoder layers. The hidden size is set to 512 while the feed-forward inner dimension is set to 1024. The multi-head attention contains 4 heads. For the larger WMT'17 dataset and the UN dataset, we adopt the configuration of *transformer_base* used in [5], which also contains 6 layers of encoder and decoder. The feed-forward inner dimension is set to 2048 and the hidden size is 512. The number of attention heads of the multi-head attention module is set to 8. The batch size is set to 4,096 source tokens and gradient accumulation of 8 is used for the WMT'17 dataset and the UN dataset to simulate multi-GPU training.

During inference, we use beam search with a beam size of 4 and length penalty of 0.6. The translation quality is evaluated with different metrics including BLEU [41] and chrF [42], [43]. We adopt the SacreBLEU toolkit[4] [44] for evaluation and test the statistical significance by bootstrap resampling [45].

C. Systems

In our experiments, we compare the following baseline methods:

1) Bilingual: We use the standard Transformer architecture [5] as our Bilingual translation baseline, in which we train an individual model for each translation direction.

2) Multilingual: We adopt the complete shared multilingual translation model [11] as our Multilingual baseline. We use the vanilla Transformer model instead of LSTM used in [11] and train a single model on different translation directions. The input sentence is prefixed with a language token to indicate the target language.

3) Linear Fusion: Our synchronous generation method with linear fusion that incorporate information from other languages with linear interpolation. The hyper-parameter λ is set to 0.1 as in previous 2-target synchronous generation methods [25].

4) Nonlinear Fusion: We use tanh as the non-linear activation function to incorporate information from other languages. Similar to *Linear Fusion*, the hyper-parameter λ is also set to 0.1 as in [25].

5) Attention Fusion: The synchronous generation with attention-based fusion that dynamically select related information from different languages.

VI. RESULTS AND ANALYSES

A. Results on the IWSLT'14 Dataset

Table III shows the translation performance of different systems on the IWSLT'14 dataset. All models are trained on

[1][Online]. Available: https://github.com/fxsjy/jieba
[2][Online]. Available: https://github.com/moses-smt/mosesdecoder
[3][Online]. Available: https://github.com/pytorch/fairseq/tree/main/examples/translation
[4][Online]. Available: https://github.com/mjpost/sacrebleu

TABLE III
THE RESULTS OF DIFFERENT METHODS ON THE IWSLT'14 DATASET. **BOLD** INDICATES BEST RESULTS OF ALL METHODS

Methods	EN→ES		EN→NL		EN→PT		EN→RO		EN→RU		EN→ZH	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
Bilingual	33.4	58.8	26.7	53.3	34.5	60.3	24.5	50.2	14.7	38.5	9.7	16.8
Multilingual	36.3	60.4	29.4	54.6	36.7	61.4	26.5	51.8	18.6	42.1	11.6	19.0
Linear Fusion	37.3	61.2	30.6	56.0	37.6	62.1	26.8	52.2	18.9	42.6	12.3	19.5
Nonlinear Fusion	37.5	61.4	30.9	56.5	39.1	63.2	27.3	53.0	18.8	42.2	12.2	19.4
Attention Fusion	**38.8**†	**62.3**†	**31.0**†	**56.5**†	**39.3**†	**63.4**†	**27.6**†	**53.3**†	**19.1**†	**42.8**†	**12.7**†	**19.9**†

† Indicate That the Corresponding Results Are Significantly Better Than the Results of Multilingual Method With $p < 0.05$.

TABLE IV
THE RESULTS ON THE UN DATASET

Methods	EN→AR		EN→ES		EN→FR		EN→RU		EN→ZH	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
Bilingual	46.9	71.7	63.5	79.1	55.9	72.6	47.3	70.4	48.5	55.9
Multilingual	45.1	70.8	62.4	78.5	55.4	72.4	45.5	69.3	47.1	54.7
Linear Fusion	45.9	71.2	62.3	78.5	55.8	72.5	46.4	69.8	47.3	54.9
Nonlinear Fusion	45.8	71.2	62.5	78.6	56.1	72.7	46.1	69.6	47.2	54.9
Attention Fusion	**46.4**†	**71.6**†	**63.3**	**79.0**†	**56.5**†	**73.0**†	**47.0**†	**70.2**†	**47.8**†	**55.4**†

Bold Indicates Best Results of Multilingual Models and the Overall Best Results are Underlined. † Indicate That the Corresponding Results are Significantly Better Than the Results of Multilingual Method With $p < 0.05$.

the extracted 7-way parallel training data. Comparing the two baseline methods, we find that the *Multilingual* method brings better translation quality for all translation directions. The reason is that the IWSLT'14 dataset is quite small (about 150 K sentences for each target language), and thus the bilingual model is not well trained. The *Multilingual* method trains a single model for all language directions and the shared parameters bring positive knowledge transfer among languages. The superiority of the *Multilingual* method suggests the effectiveness of using complementary information among languages.

We go a step further to investigate the multilingual synchronous generation that explicitly utilize the complementary information in the target side. It is obvious to see from Table III that our method performs better than the baseline methods. The *Linear Fusion* and the *Nonlinear Fusion* methods achieve similar gain over the *Multilingual* method by using a fixed amount of auxiliary information from other languages. The *Attention Fusion* dynamically selects the most relevant information from different languages and performs best among all methods, which verifies that the contribution of different languages varies in the generation process. The remarkable improvements suggest that our synchronous generation method can better explore the complementary knowledge among different languages.

B. Results on the UN Dataset

Table IV shows the translation quality on the UN dataset. Different from the results on the IWSLT'14 dataset, the *Multilingual* method performs worse than the *Bilingual* method on the UN dataset. The reason is that the dataset is quite larger than the IWSLT'14 dataset, and the bilingual model is efficiently trained. The *Multilingual* model runs into capacity bottleneck [13], [46] that a single model is not capable for modeling multiple translation directions with a massive amount of training data.

As for our method, both the *Linear Fusion* and the *Nonlinear Fusion* perform better than the *Multilingual* method by utilizing complementary knowledge of different languages, but fail in outperforming the strong *Bilingual* baseline that is efficiently trained on the large-scale bilingual data. The *Attention Fusion* method further improves the translation quality by using a more dynamic and flexible attention mechanism to incorporate information from other languages, and achieves comparable or even better performance than the strong *Bilingual* baseline.

C. Results on the WMT'17 Dataset

For the results on the WMT'17 dataset, we first evaluate the performance of our synchronous generation method trained on 4-way parallel data and report the translation quality of different systems in Table V. To make a fair comparison, the *Bilingual* and *Multilingual* models are also trained on the same 4-way training data (about 647 K parallel data). Similar to the results on the IWSLT'14 dataset, the *Multilingual* model outperforms the *Bilingual* model with the limited training data but the performance gain is less significant since the dataset is larger than the IWSLT'14 dataset. As for our method, the performance gaps between different fusion methods are smaller than the other two datasets. The reason is that there are only 3 target languages and the small number of languages limits the flexibility of the *Attention Fusion* method.

To evaluate our multi-task learning method, we use the entire training data in the WMT'17 dataset. The 4-way parallel data are used to train the synchronous generation task while the entire bilingual data are used to train the multilingual translation

TABLE V
THE RESULTS ON THE WMT'17 DATASET WITH BOTH 4-WAY PARALLEL DATA AND BILINGUAL PARALLEL DATA

Methods	Data		EN→DE		EN→LV		EN→FI	
	Bilingual	4-way Parallel	BLEU	chrF	BLEU	chrF	BLEU	chrF
Bilingual	×	✓	18.5	48.3	13.5	44.2	16.9	48.3
Multilingual	×	✓	20.4	49.3	14.4	45.0	18.5	50.1
Linear Fusion	×	✓	21.3	50.4	15.0	45.8	19.9	51.4
Nonlinear Fusion	×	✓	21.5	50.8	14.9	45.8	19.3	50.8
Attention Fusion	×	✓	21.9	51.3	15.4	46.6	20.6	52.1
Bilingual	✓	✓	28.5	56.7	17.0	46.0	22.6	54.3
Multilingual	✓	✓	27.9	55.7	18.1	48.0	23.3	55.3
Multi-Task	✓	✓	**28.4**	**56.3**†	**18.8**†	**48.8**†	**24.1**†	**56.1**†

† Indicate That the Corresponding Results Are Significantly Better Than the Results of Multilingual Method Using All Available Data With $p < 0.05$.

TABLE VI
THE COMPARISON OF MODEL SIZE AND EFFICIENCY ON THE IWSLT'14 DATASET

Model	Model Size	Efficiency			
		Training Time (h)	Training (token/s)	Generation (token/s) 1-Target	6-Target
Bilingual	287.4M	13.53	68,711.9	5,456.70	5,456.70
Multilingual	47.9M	3.81	66,845.1	5,253.33	5,253.33
Linear Fusion	47.9M	4.87	58,804.7	3,152.77	18,916.62
Nonlinear Fusion	47.9M	4.98	57,407.1	3,287.45	19,724.70
Attention Fusion	52.7M	5.39	54,483.2	2,946.97	17,681.82

The Model Size is the Number of Parameters and the Efficiency is Measured by the Number of Tokens the Model Can Process Per Second (With One NVIDIA GTX 3090 GPU) . 1-Target Means the Efficiency of Generating One Target Language at a Time While 6-Target Denotes Generating All the 6 Target Languages in the IWSLT'14 Dataset.

task as described in Section IV-D. The results are shown in Table V. With large-scale data, the advantage of *Multilingual* method is less significant, especially on EN → DE, which consists of the largest 5.9 M bilingual data. The proposed *Multi-Task* learning method, which is trained on bilingual data and multi-way parallel data with *Attention Fusion*, consistently outperforms the *Multilingual* baseline and achieves comparable or even better performance than the *Bilingual* method. Besides, compared to the *Attention Fusion* model that trained only on 4-way parallel data, the *Multi-Task* method also gains significant improvements, which proves the benefits of utilizing bilingual data when training the synchronous generation model.

D. Model Size and Efficiency

Our synchronous generation method improves the translation quality for each language with a more complicated model than the vanilla *Multilingual* method. To investigate the model size and computational efficiency of our method, we use the model trained on the IWSLT'14 dataset for evaluation and report the statistics of different systems in Table VI.

As for the model size, it is obvious that the *Bilingual* method that trains an individual model for each translation direction is the most space occupying. The *Multilingual* method trains a single model for all translation directions and the model size is much smaller for translation between multiple language pairs. Our *Linear Fusion* and *Nonlinear Fusion* methods maintain the same model size without extra parameters. The *Attention Fusion* method increases the model size with extra projection matrices in (11).

Since our method includes cross-lingual interaction with the cross-lingual attention (Section IV-B), the training and generation involve multiple languages in a single step and the computational complexity is slightly increased. As shown in the Efficiency of Table VI, our model is about 81.51% (*Attention Fusion*) to 87.97% (*Linear Fusion*) slower than the *Multilingual* baseline. In the generation for 1 target language, our method with different fusion mechanisms is much slower since they actually generate multiple target languages. When multiple languages are required (6-Target), our method achieves 3.37X (*Attention Fusion*) to 3.75X (*Nonlinear Fusion*) speedup compared to the *Multilingual* method. The statistics suggest that our method is more efficient for generating translations in multiple target languages.

E. The Contribution of Different Languages

In previous sections, we argue that our synchronous generation method can bring better translation quality by utilizing the relevant information from different languages. A natural question may arise which language is more important for translating a specific language. To answer this question, we analyze the contribution of different languages by investigating the weights of attention-based fusion in our cross-lingual attention module.

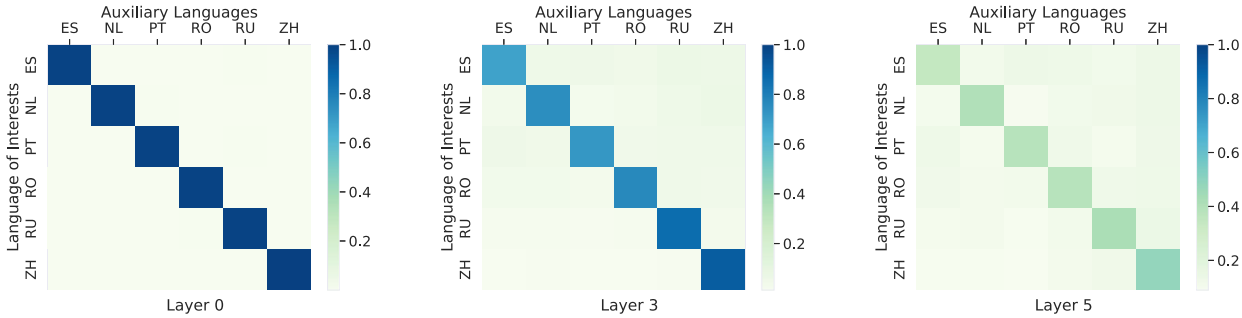

Fig. 6. Attention weights of different languages in the cross-lingual attention module with attention-based fusion in 6 decoder layers.

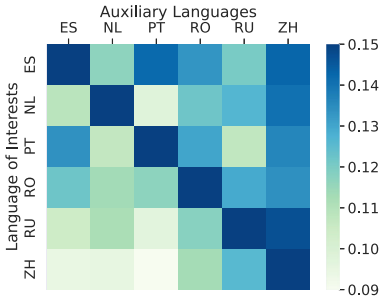

Fig. 7. The weights of cross-lingual attention with attention-based fusion in Layer 5. We set the threshold in the color map to 0.15 to better investigate the contribution of different languages.

We use the model trained on the IWSLT'14 dataset and run a forward pass over the entire validation set to obtain the cross-lingual attention weights averaged across all tokens. The attention weights of different layers are shown in Fig. 6. We find that the contribution of self-attention dynamically decreases from shallower layers to deeper layers. In the shallower layers, each language only focuses on itself and ignores the information from other languages. This may be because the representation in shallower layers only contains lexical information and the lexical knowledge among languages are quite different, e.g., the Chinese word " have the same meaning but the representations of the two words may be totally different in the shallower layers. Since the deeper layers learn more semantic level knowledge, the representation of " as the layers become deeper, and thus the complementary information from other languages can be easily exploited.

To further investigate the contribution of different languages, we plot the cross-lingual attention of the last layer (Layer 5) and set the color map threshold to 0.15. The heat map of attention weights averaged across all tokens of each sentence is shown in Fig. 7, and the attention weights at the beginning, middle, and end of each sentence are depicted in Fig. 8. In both Fig. 7 and Fig. 8, we find that Chinese is more important than other languages (the color of the last column is deeper) for prediction of different languages (different rows).

That may be because the Chinese sentences are more concise and shorter than sentences in other languages, and a word is more likely to be translated into Chinese at the early stage during generation. We also present a translation example in Table VII.

F. The Analysis of Under-Translation Errors

Since the proposed method can exploit the knowledge from both the source language and different target languages, it can potentially reduce the under-translation errors where some words in the source sentence are mistakenly untranslated. Therefore, we adopt the contrastive conditioning detection method [47] to find out under-translation errors. Specifically, given a source sentence s and a machine translation hypothesis h, we construct a set of partial source sentences s' by deleting words in s and use a pretrained mBART50 model [48] to calculate the probability score for each s' and h. For the highest probability score with a partial source s', the deleted words $s - s'$ are marked as under-translated. We refer the reader to [47] for more details.

The statistics are reported in Table VIII. It is obvious that our method can reduce the under-translation errors compared to the *Multilingual* baseline method. We also find that the rate of error reduction for different languages are inconsistent. For example, our method reduces about 28.5% under-translation errors for Dutch (NL) but achieves only 8.9% error reduction for Chinese (ZH).

G. The Effects of Different Word Order

So far, we have shown the effectiveness of our method on multilingual datasets with different scales. However, the target languages we used are mainly European languages and the word order of the languages are similar (mainly Subject-Verb-Object in terms of linguistic typology). Since our method can utilize the complementary information from multiple target languages, we also investigate the performance of our method with target languages with different word orders. We translate the English-centered parallel data of TED talks[5] collected by [49]. We choose four languages from this dataset including two Subject-Verb-Object (SVO) languages and two Subject-Object-Verb (SOV) languages. The two SVO languages are Spanish (ES) and Portuguese (PT), while the two SOV languages are Japanese (JA) and Korean (KO). We extract 277 K 5-way parallel sentences (including English) for training, 4,000 for validation and 4,000 for testing. The data processing and training settings are identical to the configuration used for the IWSLT'14 dataset.

The results measured in BLEU are shown in Table IX. Besides the *Multilingual* baseline model, we train three models with our *Attention Fusion* method to explore the effects of different word

[5][Online]. Available: https://www.ted.com

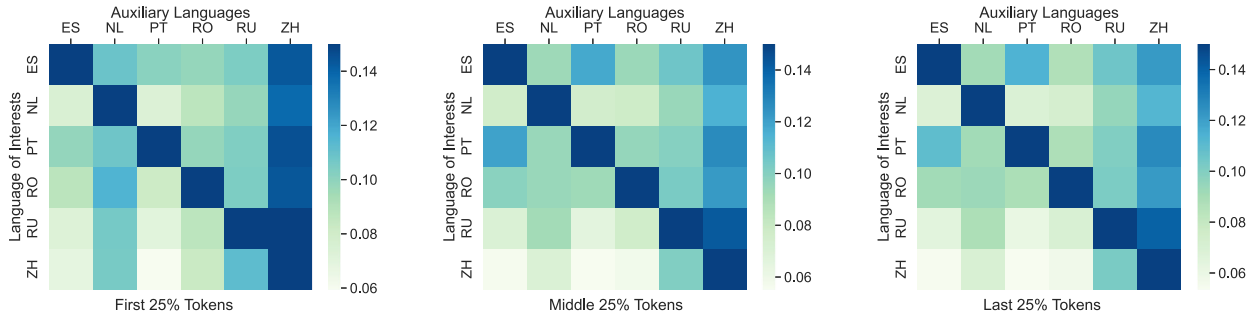

Fig. 8. The attention weights of different languages in Layer 5. We average the weights of tokens at the beginning (first 25% tokens), middle (middle 25% tokens), and end (last 25% tokens) of each sentence. The threshold in the color map is set to 0.15.

TABLE VII
TRANSLATION EXAMPLES FROM THE IWSLT'14 DATASET

Source	I asked the top scientists on this several times : Do we really have to get down to near zero ?
Reference	Le he preguntado a los más destacados científicos sobre esto varias veces , ¿ De verdad tenemos que bajar a casi cero emisiones ?
Multilingual	Le pregunté a los científicos en esto varias veces : ¿ Debemos bajar hasta casi cero ?
Our method	Le pregunté a los **destacados** científicos en esto varias veces : ¿ **De verdad** tenemos que bajar a casi cero ?
Chinese	我 问 了 很多 时候 最先 的 科学家 们 : 我们 真的 需要 降到 零 ?
Dutch	Ik vroeg de topwetenschappers vaak af : moeten we echt tot bijna nul komen ?
Source	Maybe we can't just erase 500 years of rational humanistic thought in one 18 minute speech .
Reference	Tal vez no podemos simplemente borrar 500 años de pensamiento humanístico racional en un discurso de 18 minutos .
Multilingual	Tal vez no podamos borrar 500 años de pensamiento humanista racional en 18 minutos .
Our method	Tal vez no podamos **simplemente** borrar 500 años de pensamiento humanista racional en **un discurso de** 18 minutos .
Chinese	也许 我们 不能 仅仅 在 18 分钟 的 演讲 中 删除 500 年 的 理性 的 人文主义 思维 。
Portuguese	Talvez não possamos apenas borrar 500 anos de pensamento humanista racional em um discurso de 18 minutos .

Our Method Improves the Translation in Spanish (**Bold** Words) Based on the underlined Words in Other Auxiliary Languages. All Target Language Translations Are Generated Simultaneously.

TABLE VIII
THE AVERAGE NUMBER OF UNDER-TRANSLATION ERRORS PER SENTENCE OF DIFFERENT METHODS ON THE IWSLT'14 TEST SET

	ES	NL	PT	RO	RU	ZH
Multilingual	0.159	0.159	0.118	0.026	0.215	0.264
Our Method	0.138	0.114	0.108	0.012	0.189	0.241

TABLE IX
THE PERFORMANCE OF OUR METHOD WITH TARGET LANGUAGES OF DIFFERENT WORD ORDER

	EN→ES	EN→PT	EN→JA	EN→KO
Multilingual	35.86	33.90	13.10	19.39
SVO & SOV	**37.38**	**35.28**	**14.91**	**20.62**
SVO only	36.95	34.23	-	-
SOV only	-	-	14.63	19.60

orders. The *SVO & SOV* model uses training data of English to all the four target languages, while the *SVO only* model and the *SOV only* model include only two SVO languages and two SOV languages respectively. It is obvious that the *SVO & SOV* model performs best since it can exploit more context information from multiple target languages. The results indicate that our method can benefit from more target languages with diverse word orders.

VII. CONCLUSION

In this paper, we propose a synchronous inference method for multilingual neural machine translation that allows a single model to generate translations in multiple languages simultaneously. To optimize information sharing among languages during decoding, we propose cross-lingual attention with attention-based fusion to dynamically integrate relevant information from all languages. During training the synchronous inference model, in addition to the limited multi-way parallel data, we incorporate the existing large-scale bilingual data with multi-task learning. The experimental results on three public multilingual translation datasets demonstrate that our method significantly outperforms the strong baselines and improve the decoding efficiency. Further analyses reveal that the contribution of different languages varies across layers and the contextual information of different languages are mostly fused in the deeper layers.

REFERENCES

[1] I. Sutskever *et al.*, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Systems: Annu. Conf. Neural Inf. Process. Syst.*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Montreal, QC, Canada, 2014, vol. 27, pp. 3104–3112. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[3] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, 2016, *arXiv:1609.08144*. [Online]. Available: http://arxiv.org/abs/1609.08144

[4] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. 34th Int. Conf. Mach. Learn. Res.*, D. Precup and Y. W. Teh, Eds., Sydney, NSW, Australia, 2017, vol. 70, pp. 1243–1252. [Online]. Available: http://proceedings.mlr.press/v70/gehring17a.html

[5] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Systems 30: Annu. Conf. Neural Inf. Process. Syst.*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V.N. Vishwanathan, and R. Garnett, Eds., Long Beach, CA, USA, 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[6] P. F. Brown, S. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[7] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, M. A. Hearst and M. Ostendorf, Eds., Edmonton, Canada, 2003, pp. 48–54. [Online]. Available: https://aclanthology.org/N03-1017/

[8] D. Chiang, "Hierarchical phrase-based translation," *Comput. Linguistics*, vol. 33, no. 2, pp. 201–228, 2007. [Online]. Available: https://doi.org/10.1162/coli.2007.33.2.201

[9] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-task learning for multiple language translation," in *Proc. 53rd Annu. Meeting.Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. Asian Federation Natural Lang. Process.*, Beijing, China, 2015, vol. 1, pp. 1723–1732. [Online]. Available: https://doi.org/10.3115/v1/p15-1166

[10] T.-L. Ha, J. Niehues, and A. Waibel, "Toward multilingual neural machine translation with universal encoder and decoder," in *Proc. 13th Int. Conf. Spoken Lang. Transl.*, Seattle, Washington DC, USA, 2016. [Online]. Available: https://aclanthology.org/2016.iwslt-1.6

[11] M. Johnson *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 339–351, 2017. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/1081

[12] Y. Wang, L. Zhou, J. Zhang, F. Zhai, J. Xu, and C. Zong, "A compact and language-sensitive multilingual translation method," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, L. Papers, A. Korhonen, D. R. Traum, and L. Màrquez, Eds., 2019, 2019, vol. 1, pp. 1213–1223. [Online]. Available: https://doi.org/10.18653/v1/p19-1117

[13] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, J. Su, X. Carreras, and K. Duh, Eds., Austin, TX, USA, 2016, pp. 1568–1575. [Online]. Available: https://doi.org/10.18653/v1/d16-1163

[14] O. Firat, B. Sankaran, Y. Al-Onaizan, F. T. Yarman-Vural, and K. Cho, "Zero-resource translation with multi-lingual neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, J. Su, X. Carreras, and K. Duh, Eds., Austin, Texas, USA, 2016, pp. 268–277. [Online]. Available: https://doi.org/10.18653/v1/d16-1026

[15] J. Gu, H. Hassan, J. Devlin, and V. O. K. Li, "Universal neural machine translation for extremely low resource languages," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, M. A. Walker, H. Ji, and A. Stent, Eds., 2018, vol. 1, pp. 344–354. [Online]. Available: https://doi.org/10.18653/v1/n18-1032

[16] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, K. Knight, A. Nenkova, and O. Rambow, Eds., San Diego California, USA, 2016, pp. 866–875. [Online]. Available: https://doi.org/10.18653/v1/n16-1101

[17] D. S. Sachan *et al.*, "Parameter sharing methods for multilingual self-attentional translation models," in *Proc. 3rd Conf. Mach. Translation: Research Papers*, WMT 2018, O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. L. Neves, M. Post, L. Specia, M. Turchi, and K. Verspoor, Eds., Belgium, Brussels, 2018, pp. 261–271. [Online]. Available: https://doi.org/10.18653/v1/w18-6327

[18] B. Zhang, A. Bapna, R. Sennrich, and O. Firat, "Share or not? learning to schedule language-specific capacity for multilingual translation," in *Proc. 9th Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=Wj4ODo0uyCF

[19] Y. Wang, J. Zhang, L. Zhou, Y. Liu, and C. Zong, "Synchronously generating two languages with interactive decoding," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China, 2019, pp. 3348–3353. [Online]. Available: https://doi.org/10.18653/v1/D19-1330

[20] H. He, Q. Wang, Z. Yu, Y. Zhao, J. Zhang, and C. Zong, "Synchronous interactive decoding for multilingual neural machine translation," in *Proc. 35th AAAI Conf. Artif. Intell., 33rd Conf. Innovative Appl. Artif. Intell., 11th Symp. Edu. Adv. Artif. Intell.*, 2021, pp. 12981–12988. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17535

[21] R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation," *ACM Comput. Surv.*, vol. 53, no. 5, 2020. Art. no. 99. [Online]. Available: https://doi.org/10.1145/3406095

[22] B. Zoph and K. Knight, "Multi-source neural translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, K. Knight, A. Nenkova, and O. Rambow, Eds., San Diego CA, USA, 2016, pp. 30–34. [Online]. Available: https://doi.org/10.18653/v1/n16-1004

[23] X. Zhang, J. Su, Y. Qin, Y. Liu, R. Ji, and H. Wang, "Asynchronous bidirectional decoding for neural machine translation," in *Proc. 32nd AAAI Conf. Artif. Intell. 30th Innovative Appl. Artif. Intell. 8th AAAI Symp. Edu. Adv. Artif. Intell.*, S. A. McIlraith and K. Q. Weinberger, Eds., New Orleans, LA, USA, 2018, pp. 5698–5705. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16784

[24] Z. Zheng *et al.*, "Modeling past and future for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 145–157, 2018. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/1208

[25] L. Zhou, J. Zhang, and C. Zong, "Synchronous bidirectional neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 91–105, 2019. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/1513

[26] J. Zhang, L. Zhou, Y. Zhao, and C. Zong, "Synchronous bidirectional inference for neural sequence generation," *Artif. Intell.*, vol. 281, 2020, Art. no. 103234. [Online]. Available: https://doi.org/10.1016/j.artint.2020.103234

[27] L. Zhou, J. Zhang, C. Zong, and H. Yu, "Sequence generation: From both sides to the middle," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, S. Kraus, Ed., Macao, China, 2019, pp. 5471–5477. [Online]. Available: https://doi.org/10.24963/ijcai.2019/760

[28] J. Xu and F. Yvon, "One source, two targets: Challenges and rewards of dual decoding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., 2021, pp. 8533–8546.

[29] Y. Liu *et al.*, "Synchronous speech recognition and speech-to-text translation with interactive decoding," in *Proc. 34th AAAI Conf. Artif. Intell. 32nd Innovative Appl. Artif. Intell. Conf. 10th AAAI Symp. Edu. Adv. Artif. Intell.*, New York, NY, USA, 2020, pp. 8417–8424. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/6360

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[31] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450

[32] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997. [Online]. Available: https://doi.org/10.1023/A:1007379606734

[33] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 11th IWSLT evaluation campaign," in *Proc. 11th Int. Workshop Spoken Lang. Translation: Evaluation Campaign*, M. Federico, S. Stüker, and F. Yvon, Eds., Lake Tahoe, CA, USA, 2014. [Online]. Available: https://aclanthology.org/2014.iwslt-evaluation.1

[34] M. Ziemski *et al.*, "The united nations parallel corpus v1.0," in *Proc. 10th Int. Conf. Lang. Resources Evaluation*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Portorož, Slovenia, 2016, pp. 3530–3534. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2016/summaries/1195.html

[35] O. Bojar *et al.*, "Findings of the 2017 conference on machine translation (WMT17)," in *Proc. 2nd Conf. Mach. Translation*, O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, and J. Kreutzer, Eds., Copenhagen, Denmark, 2017, pp. 169–214. [Online]. Available: https://doi.org/10.18653/v1/w17-4717

[36] P. Koehn *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proc. 45th Annu. Meeting Assoc Comput. Linguistics*, J. A. Carroll, A. van denBosch, and A.Zaenen, Eds., Prague, Czech Republic, 2007, pp. 177–180. [Online]. Available: https://aclanthology.org/P07-2045/

[37] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, vol. 1, pp. 1715–1725. [Online]. Available: https://doi.org/10.18653/v1/p16-1162

[38] M. Ott *et al.*, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Demonstrations, W. Ammar, A. Louis, and N. Mostafazadeh, Eds., Minneapolis, MN, USA, 2019, pp. 48–53.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[40] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2670313

[41] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040/

[42] M. Popovic, "chrF: Character n-gram F-score for automatic MT evaluation," in *Proc. 10th Workshop Stat. Mach. Transl., WMT, EMNLP*, Lisbon, Portugal, 2015, pp. 392–395.

[43] M. Popović, "chrF++: words helping character n-grams," in *Proc. 2nd Conf. Mach. Transl.*, Copenhagen, Denmark, O. Bojar *et al.*, Eds., 2017, pp. 612–618. [Online]. Available: https://doi.org/10.18653/v1/w17-4770

[44] M. Post *et al.*, "A call for clarity in reporting BLEU scores," in *Proc. 3rd Conf. Mach. Transl.: Res. Papers*, Belgium, Brussels, O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, C. Monz, A. Névéol, M. L. Neves, M. Post, L. Specia, M. Turchi, and K. Verspoor, Eds., 2018, pp. 186–191. [Online]. Available: https://doi.org/10.18653/v1/w18-6319

[45] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Barcelona, Spain, 2004, pp. 388–395. [Online]. Available: https://aclanthology.org/W04-3250/

[46] A. Siddhant *et al.*, "Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation," in *Proc. 34th AAAI Conf. Artif. Intell., 32nd Innov. Appl. Artif. Intell. Conf., 10th AAAI Symp. Edu. Adv. Artif. Intell.*, New York, NY, USA, 2020, pp. 8854–8861. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/6414

[47] J. Vamvas and R. Sennrich, "As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland, 2022, pp. 490–500. [Online]. Available: https://openreview.net/pdf?id=txfPhtRZ_SW

[48] Y. Tang *et al.*, "Multilingual translation from denoising pre-training," in *Proc. Findings Assoc. Comput. Linguistics: ACL/IJCNLP*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., 2021, pp. 3450–3466. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-acl.304

[49] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., 2020, pp. 4512–4525. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.365