A Compact and Language-Sensitive Multilingual Translation Method

Yining Wang[1,2], **Long Zhou**[1,2], **Jiajun Zhang**[1,2]*, **Feifei Zhai**[4], **Jingfang Xu**[4] and **Chengqing Zong**[1,2,3]
[1]National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[4]Sogou Inc., Beijing, China
{yining.wang, long.zhou, jjzhang, cqzong}@nlpr.ia.ac.cn
{zhaifeifei,xujingfang}@sogou-inc.com

Abstract

Multilingual neural machine translation (Multi-NMT) with one encoder-decoder model has made remarkable progress due to its simple deployment. However, this multilingual translation paradigm does not make full use of language commonality and parameter sharing between encoder and decoder. Furthermore, this kind of paradigm cannot outperform the individual models trained on bilingual corpus in most cases. In this paper, we propose a compact and language-sensitive method for multilingual translation. To maximize parameter sharing, we first present a universal representor to replace both encoder and decoder models. To make the representor sensitive for specific languages, we further introduce language-sensitive embedding, attention, and discriminator with the ability to enhance model performance. We verify our methods on various translation scenarios, including *one-to-many*, *many-to-many* and *zero-shot*. Extensive experiments demonstrate that our proposed methods remarkably outperform strong standard multilingual translation systems on WMT and IWSLT datasets. Moreover, we find that our model is especially helpful in low-resource and zero-shot translation scenarios.

1 Introduction

Encoder-decoder based sequence-to-sequence architecture (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Zhang and Zong, 2015; Vaswani et al., 2017; Gehring et al., 2017) facilitates the development of multilingual neural machine translation (Multi-NMT) (Dong et al., 2015; Luong et al., 2016; Firat et al., 2016; Johnson et al., 2017; Gu et al., 2018). The dominant paradigm of Multi-NMT contains one encoder to represent multiple languages and one decoder to generate output tokens of separate languages (Johnson et al., 2017; Ha et al., 2016). This paradigm is widely used in Multi-NMT systems due to simple implementation and convenient deployment.

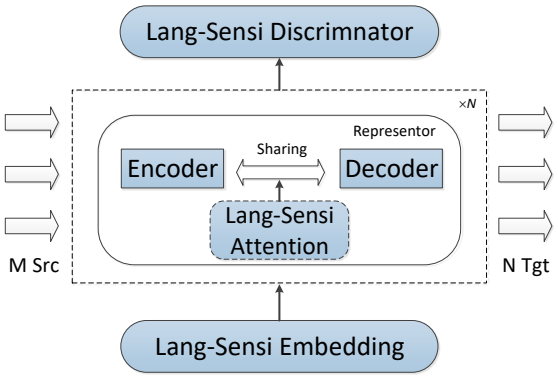

Figure 1: Our proposed compact representor, replacing encoder and decoder, can perform multilingual translation from M source languages to N target languages. We also introduce three specific modules consisting of language-sensitive embedding, language-sensitive attention, and language-sensitive discriminator.

However, this paradigm has two drawbacks. For one hand, using single encoder-decoder framework for all language pairs usually yields inferior performance compared to individually trained single-pair models in most cases (Lu et al., 2018; Platanios et al., 2018; Wang et al., 2018). For the other hand, although this paradigm saves lots of parameters compared to another Multi-NMT framework which employs separate encoders and decoders to handle different languages (Dong et al., 2015; Zoph and Knight, 2016; Luong et al., 2016; Firat et al., 2016), parameter sharing between encoder and decoder are not fully explored. Since both encoder and decoder have similar

* Jiajun Zhang is the corresponding author and the work is done while Yining Wang is doing research intern at Sogou Inc.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1213–1223
Florence, Italy, July 28 - August 2, 2019.

structures but use different parameters, the commonality of languages cannot be fully exploited in this paradigm. A natural question arises that why not share the parameters between encoder and decoder on multilingual translation scenario?

To address these issues, we present a compact and language-sensitive method in this work, as shown in Figure 1. We first propose a unified representor by tying encoder and decoder weights in Multi-NMT model, which can not only reduce parameters but also make full use of language commonality and universal representation. To enhance the model ability to distinguish different languages, we further introduce language-sensitive embedding, attention, and discriminator.

We conduct extensive experiments to verify the effectiveness of our proposed model on various Multi-NMT tasks including one-to-many and many-to-many which is further divided into balanced, unbalanced and zero-shot. Experimental results demonstrate that our model can significantly outperform the strong standard baseline multilingual systems and achieve even better performance than individually trained models on most of the language pairs.

Specifically, our contributions are three-fold in this work:

(1) We present a universal representor to replace encoder and decoder, leading to a compact translation model, which fully explores the commonality between languages.

(2) We introduce language-sensitive embedding, attention, and discriminator which augment the ability of Multi-NMT model in distinguishing different languages.

(3) Extensive experiments demonstrate the superiority of our proposed method on various translation tasks including one-to-many, many-to-many and zero-shot scenarios. Moreover, for many-to-many using unbalance translation pairs, we can achieve the new state-of-the-art results on IWSLT-15 English-Vietnamese. For zero-shot translation, our methods can achieve even better results than individually trained models with the parallel corpus.

2 Background

In this section, we will introduce the background of the encoder-decoder (Sutskever et al., 2014; Cho et al., 2014) framework and self-attention-based Transformer (Vaswani et al., 2017).

2.1 Encoder-Decoder Framework

Given a set of sentence pairs $D = \{(\mathbf{x}, \mathbf{y})\}$, the encoder f_{enc} with parameters θ_{enc} maps an input sequence $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ to a sequence of continuous representations $h^{\text{enc}} = (h_1^{\text{enc}}, h_2^{\text{enc}}, \cdots, h_n^{\text{enc}})$ whose size varies concerning the source sentence length. The decoder f_{dec} with θ_{dec} generates an output sequence $\mathbf{y} = (y_1, y_2, \cdots, y_m)$ by computing $P(y_t|y_{<t})$ as follows:

$$P(y_t|y_{<t}) = \text{softmax}(f(h^{\text{dec}}, c_t)) \quad (1)$$

where h^{dec} is a sequence of continuous representations for the decoder and c_t is the context vector which can be calculated as follows:

$$c_t = \sum_{i=1}^{n} a_{i,t} h_i^{\text{enc}} \quad (2)$$

where $a_{i,t}$ is attention weight:

$$a_{i,t} = \text{softmax}(e_{i,t}) = \frac{\exp e_{i,t}}{\sum_{j=1}^{n} \exp e_{j,t}} \quad (3)$$

where $e_{i,t}$ is a similarity score between the source and target representations. The parameters of calculating cross-attention weight $a_{i,t}$ are denoted as θ_{attn}.

The encoder and decoder are trained to maximize the conditional probability of target sequence given a source sequence:

$$\mathcal{L}_t(D; \boldsymbol{\theta}) = \sum_{d=1}^{|D|} \sum_{t=1}^{M} \log P(y_t|y_{<t}, x; \theta_{\text{enc}}, \theta_{\text{dec}}, \theta_{\text{attn}}) \quad (4)$$

where M is target sentence length. For simplicity, we do not specify d in this formula.

Both the encoder and decoder can be implemented by the different basic neural models structures, such as RNN (LSTM/GRU) (Sutskever et al., 2014; Cho et al., 2014), CNN (Gehring et al., 2017), and self-attention (Vaswani et al., 2017). Our proposed method can be applied to any encoder-decoder architecture. Considering the excellent translation performance of self-attention based Transformer (Vaswani et al., 2017), we implement our method based on this architecture.

2.2 Transformer Network

Transformer is a stacked network with several layers containing two or three basic blocks in each layer. For a single layer in the encoder, it consists of a multi-head self-attention and a position-wise

feed-forward network. For the decoder model, besides the above two basic blocks, a multi-head cross-attention follows multi-head self-attention. In this block, the calculation method of similarity score e_t in Equation 3 is a little different from Luong et al. (2015) and Bahdanau et al. (2015):

$$e_{i,t} = \frac{1}{\sqrt{d_m}} W_k h_i^{\text{enc}} * W_q h_t^{\text{dec}} \quad (5)$$

where d_m is the dimension of hidden units, W_k and W_q are parameters of this cross-attention block, which are denoted as θ_{attn} in Equation 4.

All the basic blocks are associated with residual connections, followed by layer normalization (Ba et al., 2016). Since the Transformer network contains no recurrence, positional embeddings are used in the model to make use of sequence order. More details regarding the architecture can be found in Vaswani et al. (2017).

2.3 Multilingual Translation

In contrast to NMT models, multilingual models perform the multi-task paradigm with some degree of parameter sharing, in which models are jointly trained on multiple language pairs. We mainly focus on mainstream multilingual translation method proposed by Johnson et al. (2017), which has a unified encoder-decoder framework with a shared attention module for multiple language pairs. They decompose the probability of the target sequences into the products of per token probabilities in all translation forms:

$$\mathcal{L}_{m-t}(D;\boldsymbol{\theta}) = \sum_{l=1}^{L}\sum_{d=1}^{|D_l|}\sum_{t=1}^{M} \log P(y_t^l|x^l, y_{<t}^l; \theta_{\text{enc}}, \theta_{\text{dec}}, \theta_{\text{attn}}) \quad (6)$$

where L is the number of translation pairs and $P(y_t^l|x^l, y_{<t}^l; \boldsymbol{\theta})$ denotes the translation probability of t-th word of the d-th sentence in l-th translation pair. Note that the translation process for all target languages uses the same parameter set $\boldsymbol{\theta}$.

3 Our Method

In this section, we introduce our compact and language-sensitive method for multilingual translation, which can compress the model by a representor and improve model ability with language-sensitive modules.

3.1 A Compact Representor

In Multi-NMT model, the encoder and decoder are two key components, which play analogous roles and have a similar structure in each layer. We argue that encoder and decoder can share the same parameters if necessary. Thus, we introduce a representor to replace both encoder and decoder by sharing weight parameters of the self-attention block, feed-forward block and the normalization block, as shown in Figure 2. The representor parameters are denoted θ_{rep}. Therefore, the objective function (Equation 6) becomes:

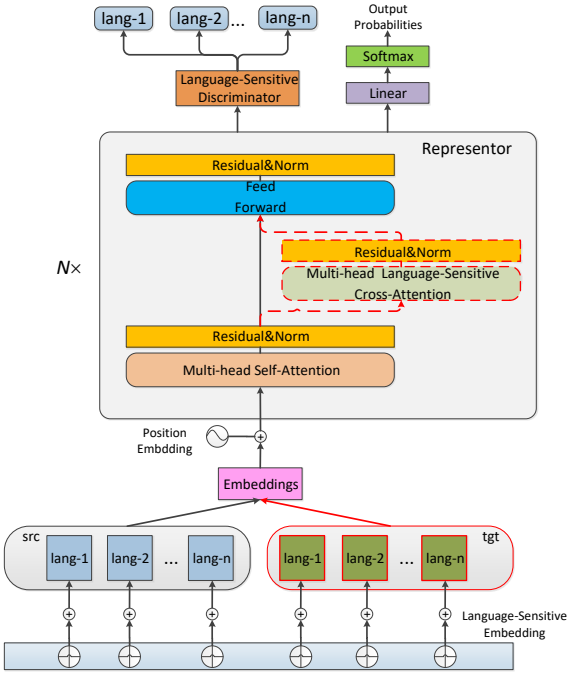

Figure 2: The framework of Multi-NMT using our compact and language-sensitive method.

$$\mathcal{L}_{m-t}(D;\boldsymbol{\theta}) = \sum_{l=1}^{L}\sum_{d=1}^{|D_l|}\sum_{t=1}^{M} \log P(y_t^l|x^l, y_{<t}^l; \theta_{\text{rep}}, \theta_{\text{attn}}) \quad (7)$$

This representor (θ_{rep}) coordinates the semantic presentation of multiple languages in a closely related universal level, which also increases the utilization of commonality for different languages.

3.2 Language-Sensitive Modules

The compact representor maximizes the sharing of parameters and makes full use of language commonality. However, it lacks the ability to discriminate different languages. In our method, we introduce three language-sensitive modules to enhance our model as follows:

1) **Language-Sensitive Embedding**: Previously, Press and Wolf (2017) conduct the weight tying of input and output embedding in NMT model. Generally, a shared vocabulary is built

upon subword units like BPE (Sennrich et al., 2016b) and wordpiece (Wu et al., 2016; Schuster and Nakajima, 2012). However, it remains under-exploited which kind of embedding sharing is best for Multi-NMT. We divide the sharing manners into four categories including *language-based* manner (**LB**, different languages have separate input embeddings), *direction-based* manner (**DB**, languages in source side and target side have different input embeddings), *representor-based* manner (**RB**, shared input embeddings for all languages) and *three-way weight tying* manner (**TWWT**) proposed in Press and Wolf (2017), in which the output embedding of the target side is also shared besides representor-based sharing. We compare these four sharing manners for Multi-NMT in our experiments, and we will discuss the results in Section 5.

Considering the last three sharing manners cannot model a sense of which language a token belongs to, we propose a new language-sensitive embedding in our method to specify different languages explicitly. Similar to the position embeddings described in Section 2, this kind of embedding is added to the embedding of each token for corresponding language, which can indicate the translation direction on the source side and guide the generation process for target languages. This embedding is denoted as $E_{\text{lang}} \in \mathbb{R}^{|K| * d_{\text{model}}}$, where $|K|$ is the number of languages involved, and d_{model} is the dimension of hidden states in our model. Note that this embedding can be learned during training.

2) **Language-Sensitive Attention**: In NMT architecture, cross-attention only appearing in the decoder network locates the most-relevant source part when generating each token in target language. For Multi-NMT, we introduce three different ways to design the cross-attention mechanism, consisting of **i)** *shared-attention*, **ii)** *hybrid-attention*, and **iii)**) *language-sensitive attention* utilized in our method.

i): In our proposed compact representor, we share self-attention block between encoder and decoder. For the *shared-attention*, we make a further step to share parameters of cross-attention and self-attention, which can be regarded as coordination of information from both the source side and target side.

ii): Different from the above attention mechanism, the *hybrid-attention* utilizes independent cross-attention modules but it is shared for all translation tasks.

iii): In the *language-sensitive attention*, it allows the model to select the cross-attention parameters associated with specific translation tasks dynamically.

In our paper, we investigate these three attention mechanisms. We argue that both the shared and hybrid mechanisms tend to be confused to extract information from different source languages when decoding multiple source languages with different word orders. Thus, we mainly focus on languages-sensitive attention in our method. To this end, we use multiple sets of parameters θ_{attn} to represent cross-attention modules of different translation tasks. However, *language-sensitive attention* does not support zero-shot translation because there is no explicit training set for this specific translation task. Therefore, we employ *hybrid-attention* mechanism in our zero-shot experiments.

3) **Language-Sensitive Discriminator**: In our method, the representor which shares encoder and decoder makes full use of language commonality, but it weakens the model ability to distinguish different languages. Hence we introduce a new language-sensitive discriminator to strengthen model representation.

In NMT framework, the hidden states on the top layer can be viewed as a fine-grained abstraction (Anastasopoulos and Chiang, 2018). For this language-sensitive module, we first employ a neural model f_{dis} on the top layer of reprensentor $h^{\text{rep}}_{\text{top}}$, and the output of this model is a language judgment score P_{lang}.

$$\begin{aligned} h^{\text{dis}} &= f_{\text{dis}}(h^{\text{rep}}_{\text{top}}) \\ P_{\text{lang}}(d) &= \text{softmax}(W_{\text{dis}} * h^{\text{dis}}_{d} + b_{\text{dis}}) \end{aligned} \tag{8}$$

where $P_{\text{lang}}(d)$ is language judgment score for sentence pair d, W_{dis}, b_{dis} are parameters, which are denoted as θ_{dis}. We test two different types of neural models for f_{dis}, including convolutional network with max pooling layer and two-layer feed-forward network.

And then, we obtain an discriminant objective function as follows:

$$\mathcal{L}_{\text{dis}}(\theta_{\text{dis}}) = \sum_{k \in K} \sum_{d=1}^{|D|} \mathbb{I}\{g_d = k\} * \log P_{\text{lang}}(d) \tag{9}$$

where $\mathbb{I}\{\cdot\}$ is indicator function, and g_d belongs to language k.

Finally, we incorporate the language-sensitive discriminator into our Multi-NMT model, and it can be optimized through an end-to-end manner for all translation language pairs D with the following objective function.

$$\begin{aligned}\mathcal{L}(D;\boldsymbol{\theta}) =& \mathcal{L}(D;\theta_{\text{rep}},\theta_{\text{attn}},\theta_{\text{dis}}) \\ =&(1-\lambda)\mathcal{L}_{m-t}(\theta_{\text{rep}},\theta_{\text{attn}}) + \lambda\mathcal{L}_{\text{dis}}(\theta_{\text{dis}})\end{aligned} \tag{10}$$

where λ is learned or pre-defined weight to balance the translation task and language judgment task.

4 Experimental Settings

4.1 Data

In this section, we describe the datasets using in our experiments on one-to-many and many-to-many multilingual translation scenarios.

One-to-Many: For this translation scenario, we perform one-to-two, one-to-three, and one-to-four multilingual translation on the combination of WMT-14[1] (English-to-German, briefly En→De), WMT-17[2] datasets (English-to-Latvian, briefly En→Lv) and WMT-18[3] (English-to-Finnish, English-to-Chinese without UN part[4], briefly En→Fi and En→Zh) datasets.

Many-to-Many: For many-to-many translation, we test our methods on IWSLT-17[5] translation datasets, including English, Italian, Romanian, Dutch (briefly, En, It, Ro, Nl). In order to perform zero-shot translation, we discard some particular language pairs. We also evaluate our method on the unbalanced training corpus. To this end, we construct the training corpus using resource-rich En-De, En-Fi in WMT datasets and low-resource English-Vietnamese (briefly, En-Vi) in IWSLT-15[6].

The statistical information of all the datasets is detailed in Table 1.

Datasets	Language pair	Train	Dev	Test
WMT	En-De	4.50M	6003	3003
	En-Lv	4.50M	2003	2001
	En-Fi	3.25M	3000	3000
	En-Zh	9.02M	2002	2001
IWSLT	En-It	231.6k	929	1566
	En-Ro	220.5k	914	1678
	En-Nl	237.2k	1003	1777
	Ro-It	217.5k	914	1643
	En-Vi	130.9k	768	1268

Table 1: The statistics of all the datasets including WMT and IWSLT tasks.

4.2 Training Details

We implement our compact and language-sensitive method for Multi-NMT based on the tensor2tensor[7] library. We use wordpiece method (Wu et al., 2016; Schuster and Nakajima, 2012) to encode the combination of both source side sentences and target side sentences. The vocabulary size is 37,000 for both sides. We train our models using configuration *transformer_base* adopted by Vaswani et al. (2017), which contains a 6-layer encoder and a 6-layer decoder with 512-dimensional hidden representations. Each mini-batch contains roughly 3,072 source and 3,072 target tokens, which belongs to one translation direction. We use Adam optimizer (Kingma and Ba, 2014) with β_1=0.9, β_2=0.98, and ϵ=10^{-9}. For evaluation, we use beam search with a beam size of $k = 4$ and length penalty $\alpha = 0.6$. All our methods are trained and tested on a single Nvidia P40 GPU.

5 Results and Analysis

In this section, we discuss the results of our experiments about our compact and language-sensitive method on Multi-NMT. The translation performance is evaluated by character-level BLEU5 for En→Zh translation and case-sensitive BLEU4 (Papineni et al., 2002) for other translation tasks. In our experiments, the models trained on individual language pair are denoted by *NMT Baselines*, and the baseline Multi-NMT models are denoted by *Multi-NMT Baselines*.

5.1 One-to-Many Translation

5.1.1 Main Results

The main results on the one-to-many translation scenario, including one-to-two, one-to-three and one-to-four translation tasks are reported in Table 2. We present a typical Multi-NMT adopting Johnson et al. (2017) method on Transformer as our *Multi-NMT baselines* model. Obviously, *Multi-NMT Baselines* cannot outperform *NMT Baselines* in all cases, among which four directions are comparable and twelve are worse.

[1] http://www.statmt.org/wmt14/translation-task.html
[2] http://www.statmt.org/wmt17/translation-task.html
[3] http://www.statmt.org/wmt18/translation-task.html
[4] https://cms.unov.org/UNCorpus/
[5] https://sites.google.com/site/iwsltevaluation2017
[6] https://sites.google.com/site/iwsltevaluation2015
[7] https://github.com/tensorflow/tensor2tensor

Task	Tgt	*NMT Baselines*	*Multi-NMT Baselines* Johnson et al. (2017)	*Three-Stgy* Wang et al. (2018)	*Rep+Emb*	*Rep+Emb +Attn*	*Rep+Emb +Attn+Dis*
One-to-Two	De	27.50	27.26	27.35	26.60	26.96	**27.74**
	Lv	16.28	16.32	16.38	15.37	15.87	**16.79**
	De	27.50	27.88	27.89	26.96	27.32	**27.96**
	Fi	16.83	16.47	16.70	15.78	16.58	**16.89**
	De	**27.50**	26.80	26.99	26.08	26.68	27.45
	Zh	26.04	25.54	25.78	24.48	25.33	**26.17**
One-to-Three	De	**27.50**	25.44	25.55	24.82	25.45	26.06
	Zh	26.04	24.87	25.63	24.12	24.93	**26.12**
	Fi	16.83	16.86	16.97	16.06	16.78	**17.12**
	De	**27.50**	25.98	26.12	24.88	25.80	26.42
	Lv	16.28	14.88	15.44	14.51	15.58	**16.31**
	Fi	16.83	16.94	17.05	16.15	16.79	**17.22**
One-to-Four	De	**27.50**	23.59	22.88	22.88	23.58	24.08
	Lv	16.28	15.57	16.02	15.00	16.21	**16.57**
	Zh	26.04	25.24	25.83	24.15	25.27	**26.29**
	Fi	**16.83**	13.45	14.12	12.99	14.11	15.03

Table 2: Translation performance on one-to-two, one-to-three and one-to-four translation tasks. *Rep* denotes our proposed representor. *Emb*, *Attn*, and *Dis* represent our proposed language-sensitive methods to address multilingual translation. Note that the source language of all our experiments is English.

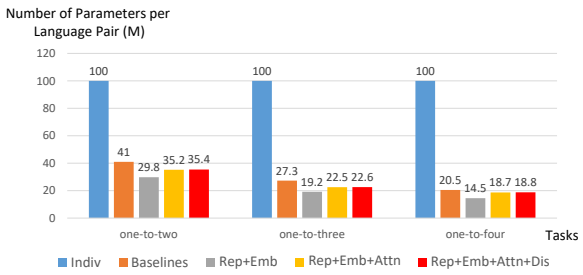

Figure 3: The comparison of model scale among individually trained system, baselines Multi-NMT system and our methods. Y-axis represents the model parameters per language pair, which is calculated by averaging model parameters on all translation tasks involved.

Src→Tgt	Emb Manners	Size	Tgt-1	Tgt-2
En→De/Lv	**LB**	139M	26.58	15.76
	DB	100M	27.22	16.26
	RB	82M	**27.26**	**16.32**
	TWWT	63M	26.82	16.02
En→De/Zh	**LB**	139M	**27.34**	**25.61**
	DB	100M	27.15	25.22
	RB	82M	27.22	25.38
	TWWT	63M	26.91	24.99

Table 3: Size (number of parameters) and BLEU scores of various embedding sharing manners. **LB**, **DB**, **RB**, **TWWT** denote language-based manner, direction-based manner, representor-based manner, and three-way weight tying manner separately, as mentioned in Section 3.2. Tgt-1 and Tgt-2 mean the results of the first (De) and the second (Lv/Zh) target language.

With respect to our proposed method, it is clear that our compact method consistently outperforms the baseline systems. Compared with another strong one-to-many translation model *Three-Stgy* proposed by Wang et al. (2018), our compact method can achieve better results as well. Moreover, our method can perform even better than individually trained systems in most cases (eleven out of sixteen cases). The results demonstrate the effectiveness of our method.

5.1.2 Model Size

Besides improving the translation results, we also compress the model size by introducing the representor. We investigate the scale of parameters used on average in each translation direction. We compare three models, including *NMT Baselines* model, *Multi-NMT Baselines* model, and our compact Multi-NMT model. As shown in Figure 3, all the multilingual translation models reduce the parameters. Compared with *Multi-NMT Baselines*, we can observe that our method further reduces the model size of Multi-NMT. Considering Table 2 and Figure 3 together, we note that even though our proposed method in one-to-four translation task only uses 18.8% parameters of *NMT Baselines*, we can achieve better performance on En→Zh and En→Lv.

5.1.3 Discussion of Language-Sensitive Modules

Table 2 shows that our proposed language-sensitive modules are complementary with each other. In this subsection, we will analyze each module in detail.

Language-Sensitive Embedding: As mentioned in section 3.2, embedding sharing man-

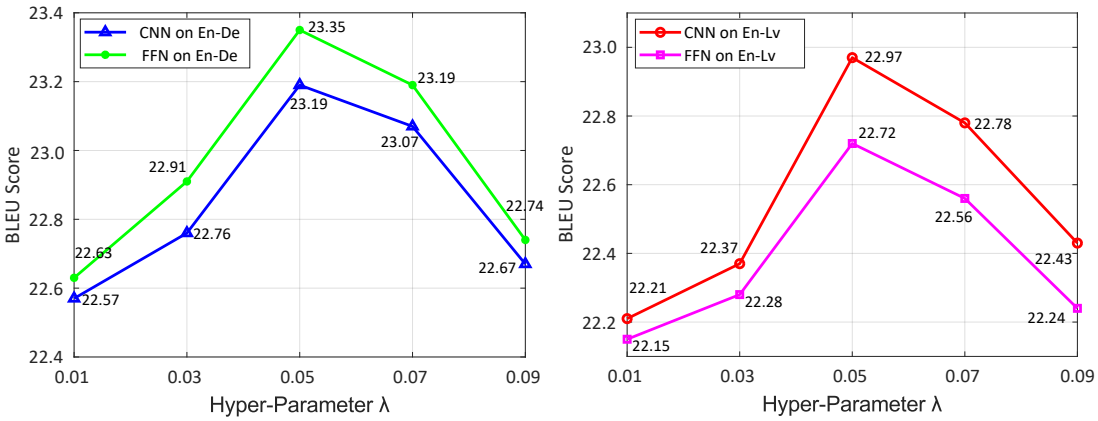

Figure 4: The comparison of two neural models with different hyper-parameter λ. CNN and FFN denote convolution network and feed-forward network, respectively.

ners for Multi-NMT are divided into four categories. We show the results of these sharing manners in Table 3. To make a fair comparison, we sample 4.5M sentence pairs from En-Zh dataset. As shown in this table, our representor-based sharing manner consistently outperforms both the direction-based manner and three-way weight tying manner. Furthermore, even though the representor-based manner has about 40% fewer parameters than the language-based manner, it achieves comparable or even better performance. We find that language-based sharing manner is unstable because it achieves the highest BLEU score on Multi-NMT of similar languages (En→De/Zh), but the worst quality on dissimilar languages (En→De/Lv). Taking into account of translation quality and stability, we choose to use representor-based sharing manner in our method.

As described in Section 3.2, our proposed language-sensitive embedding is added to the input embedding of each token, which is unlike convention Multi-NMT method adding a special token into source side sentences or vocabularies (Johnson et al., 2017; Ha et al., 2016). There exists a question, is this kind of embeddings essential in our representor? To make a verification, we do the ablation study without this module. We observe that Multi-NMT model does not converge during training, which demonstrates these language-sensitive embeddings play a significant role in our model.

Language-Sensitive Attention: We present three types of cross-attention mechanisms in Section 3.2. We adopt *shared-attention* and *language-sensitive attention* for *Rep+Emb* and *Rep+Emb+Attn* separately. Comparing these two methods in Table 2, *Rep+Emb+Attn* method outperforms *Rep+Emb* method in all cases, which demonstrates the language-sensitive is useful for multiple language pairs with different word order. We also conduct the experiment of our representor with the *hybrid-attention* mechanism. Since this method has similar performance with *Rep+Emb* but is larger in size, we ignore its results here.

Language-Sensitive Discriminator: In section 3.2, we employ two different types of the neural model as a language-sensitive discriminator, and there is a hyper-parameter λ in Equation 10. We present the effect of convolutional network and feed-forward network with different hyper-parameters on development datasets in Figure 4. Considering that distinguishing between languages is only an auxiliary task in Multi-NMT, we set the maximum of λ to be 0.1. As shown in Figure 4, when we adopt the convolution network as our discriminator with $\lambda = 0.05$, our language-sensitive method performs best. We also conduct the experiments in which the hyper-parameter λ is learnable. The experiment results are similar to the best settings mentioned above both on En→De (23.35 vs. 23.19) and En→Lv (22.97 vs. 22.72). For simplicity, all our experiments listed in Table 2 and 4 adopt convolution network as the language-sensitive discriminator with $\lambda = 0.05$.

5.2 Many-to-Many Translation

Table 4 reports the detailed results of different methods under the many-to-many translation scenario. We will analyze the performance below.

Task	Src→Tgt	*NMT Baselines*	*Multi-NMT Baselines* Johnson et al. (2017)	*Rep+Emb*	*Rep+Emb +Attn*	*Rep+Emb +Attn+Dis*
Many-to-Many for Balanced Corpus						
I Supervised Four-to-Four	En→It	28.41	29.53	29.47	29.98	**30.23**
	It→En	30.66	31.70	31.76	32.23	**32.75**
	En→Ro	21.41	22.23	22.16	22.87	**23.53**
	Ro→En	26.09	27.69	27.58	27.98	**28.32**
	En→Nl	25.88	27.88	26.96	27.32	**27.96**
	Nl→En	27.48	28.67	28.58	28.86	**29.32**
	It→Ro	12.77	13.86	13.89	14.35	**14.89**
	Ro→It	13.54	14.78	14.66	14.87	**15.22**
II Zero-Shot	Nl→Ro	14.15	13.70	13.98	15.12	**15.54**
	Ro→Nl	14.33	13.91	14.17	14.86	**15.41**
	It→Nl	18.24	17.97	18.02	18.98	**19.74**
	Nl→It	18.11	17.59	18.16	19.18	**19.87**
Many-to-Many for Unbalanced Corpus						
III Supervised Three-to-Three	En→De	**27.60**	24.39	23.78	25.45	26.06
	De→En	**32.23**	28.85	28.14	28.98	30.37
	En→Fi	**16.83**	14.58	13.82	14.26	14.77
	Fi→En	**22.37**	19.60	19.15	19.96	21.03
	En→Vi	26.78	28.89	28.84	30.49	**32.01**
	Vi→En	25.72	27.19	27.27	29.14	**31.71**

Table 4: Translation performance under the many-to-many scenario, consisting of supervised four-to-four and zero-shot translation on the balanced corpus, and supervised three-to-three on the unbalanced corpus. Note that we do not use the Nl-Ro and It-Nl language pairs in our many-to-many translation task for the balanced corpus.

5.2.1 Results of Balanced Corpus

In part I of Table 4, our compact and language-sensitive method (*Rep+Emb+Attn+Dis*) performs consistently better than corresponding *Multi-NMT Baselines*, and it can achieve the improvements up to 1.30 BELU points (23.53 vs. 22.23 on En→Ro). Although *Rep+Emb* method dramatically reduces the model parameters, it performs on par with *Multi-NMT Baselines*. Compared with *NMT Baselines* model, our method also achieves better results, which is nearly 2 BLEU points on average. Experimental results on our balanced corpus demonstrate that our method is robust and valid under the many-to-many translation scenario.

5.2.2 Results of Unbalanced Corpus

For unbalanced corpus, our method can achieve better results than *Multi-NMT Baselines* as well, as shown in part III of Table 4. Moreover, from the last two lines of this part, we can observe that compared with *NMT Baselines*, the translation quality of En↔Vi can achieve the improvements up to 5.23/5.99 BLEU points (32.01/31.71 vs. 26.78/25.72), both of which are new state-of-the-art on these translation tasks to the best of our knowledge. The results show that our method is more effective in low-resource language pairs, especially for the unbalanced corpus.

5.2.3 Zero-Shot Results

Part II in Table 4 shows the performance of zero-shot translation. Note that we conduct experiments of this translation scenario using *hybrid-attention* mechanism. Compared with *Multi-NMT Baselines*, our compact and language-sensitive method performs significantly better with the improvement as large as 2.28 BLEU points on Nl→It. Note that the training datasets do not contain parallel data for Nl-Ro and It-Nl.

It is interesting to figure out the translation performance of Nl↔Ro and It↔Nl when bilingual training corpus is available. We conduct experiments of *NMT Baselines* on Nl-Ro and It-Nl with all sentence pairs in IWSLT-17 (about 200k), which is similar to other training pairs in our balanced corpus. As shown in part II, *Multi-NMT Baselines* underperform the *NMT Baselines* on all cases. However, our method performs better than *NMT Baselines*, and it achieves the improvement up to 1.76 BLEU points on Nl→It translation task.

6 Related Work

Our work is related to two lines of research, and we describe each of them as follows:

Model Compactness and Multi-NMT: To reduce the model size in NMT, weight pruning, knowledge distillation, quantization, and weight sharing (Kim and Rush, 2016; See et al., 2016; He et al., 2018; Zhou et al., 2018) have been ex-

plored. Due to the benefit of compactness, multilingual translation has been extensively studied in Dong et al. (2015), Luong et al. (2016) and Johnson et al. (2017). Owing to excellent translation performance and ease of use, many researchers (Blackwood et al., 2018; Lakew et al., 2018) have conducted translation based on the framework of Johnson et al. (2017) and Ha et al. (2016). Zhou et al. (2019) propose to perform decoding in two translation directions synchronously, which can be applied on different target languages and is a new research area for Multi-NMT. In our method, we present a compact method for Multi-NMT, which can not only compress the model but also yield superior performance.

Low-Resource and Zero-Shot NMT: Many researchers have explored low-resource NMT using transfer learning (Zoph et al., 2016; Neubig and Hu, 2018) and data augmenting (Sennrich et al., 2016a; Zhang and Zong, 2016) approaches. For zero-shot translation, Cheng et al. (2017) and Chen et al. (2017) utilize a pivot-based method, which bridges the gap between source-to-pivot and pivot-to-target two steps. Multilingual translation is another direction to deal with both low-resource and zero-shot translation. Gu et al. (2018) enable sharing of lexical and sentence representation across multiple languages, especially for extremely low-resource Multi-NMT. Firat et al. (2016), Lakew et al. (2017), and Johnson et al. (2017) propose to make use of multilinguality in Multi-NMT to address the zero-shot problem. In this work, we propose a method for Multi-NMT to boost the accuracy of the multilingual translation, which better fits on both low-resource scenario and zero-shot scenario.

7 Conclusion

In this paper, we have proposed a compact and language-sensitive method for multilingual translation. We first introduce a representor for replacing both encoder and decoder so as to fully explore the commonality among languages. Based on the representor architecture, we then propose three language-specific modules dealing with embedding, attention and language discrimination respectively, in order to enhance the multilanguage translation model with the ability of distinguishing among different languages. The empirical experiments demonstrate that our proposed methods can outperform strong standard multilingual translation systems on one-to-many and many-to-many translation tasks. Moreover, our method is proved to be especially helpful in the low-resource and zero-shot translation scenarios.

Acknowledgments

The research work descried in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2016QY02D0303 and the Natural Science Foundation of China under Grant No. U1836221 and 61673380. The research work in this paper has also been supported by Beijing Advanced Innovation Center for Language Resources and Sogou Inc. We would like to thank Yang Zhao and Yuchen Liu for their invaluable discussions on this paper.

References

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. *In Proceedings of NAACL 2018*, pages 82–91.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *In Proceedings of ICLR 2015*.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. *In Proceedings of COLING 2018*, pages 3112–3122.

Yun Chen, Yong Cheng, Yang Liu, and Li Victor, O.K. 2017. A teacher-student framework for zero-resource neural machine translation. *In Proceedings of ACL 2017*, pages 1925–1935.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. *Proceedings of IJCAI 2017*, pages 3974–3980.

Kyunghyun Cho, Bart van Merriënboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *In Proceedings of EMNLP 2014*, pages 1724–1734.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. *In Proceedings of ACL 2015*, pages 1723–1732.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *In Proceedings of NAACL 2016*, pages 866–875.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1601.03317*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. *In Proceedings of NAACL 2018*, pages 344–354.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *In Proceedings of IWSLT 2016*.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. *In Proceedings of NIPS 2018*, pages 7944–7954.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. *In Proceedings of EMNLP 2013*, pages 1700–1709.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. *In Proceedings of EMNLP 2016*, pages 1317–1327.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Surafel M Lakew, ADG Mattia, and F Marcello. 2017. Multilingual neural machine translation for low resource languages. *CLiC-it*.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. *In Proceedings of COLING 2018*, pages 641–652.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. *In Proceedings of WMT 2018*, pages 84–92.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. *In Proceedings of ICLR 2016*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *In Proceedings of EMNLP*, pages 1412–1421.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *In Proceedings of EMNLP 2018*, pages 875–880.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *In Proceedings of ACL*, pages 311–318.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. *In Proceedings of EMNLP 2018*, pages 425–435.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. *In Proceedings of EACL 2017*, pages 157–163.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. *In Proceedings of ICASSP 2012*.

Abigail See, Minh-Thang Luong, and Christopher D. Manning. 2016. Compression of neural machine translation models via pruning. *In Proceedings of SIGNLL 2016*, pages 291–301.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. *In Proceedings of ACL 2016*, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. *In Proceedings of ACL 2016*, pages 1715–1725.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *In Proceedings of NIPS*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Łukasz Kaiser. 2017. Attention is all you need. *In Proceedings of NIPS*, pages 30–34.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. *In Proceedings of EMNLP 2018*, pages 2955–2960.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jiajun Zhang and Chengqing Zong. 2015. Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, 30(5):16–25.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. *In Proceedings of EMNLP*, pages 1535–1545.

Long Zhou, Yuchen Liu, Jiajun Zhang, Chengqing Zong, and Guoping Huang. 2018. Language-independent representor for neural machine translation. *arXiv preprint arXiv:1811.00258*.

Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019. Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics*, 7:91–105.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *In Proceedings of NAACL 2016*, pages 30–34.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *In Proceedings of EMNLP 2016*, pages 1568–1575.