Synchronously Generating Two Languages with Interactive Decoding

Yining Wang[1,2], **Jiajun Zhang**[1,2]*, **Long Zhou**[1,2]
Yuchen Liu[1,2] 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
{yining.wang, jjzhang, long.zhou}@nlpr.ia.ac.cn
{yuchen.liu, cqzong}@nlpr.ia.ac.cn

Abstract

In this paper, we introduce a novel interactive approach to translate a source language into two different languages simultaneously and interactively. Specifically, the generation of one language relies on not only previously generated outputs by itself, but also the outputs predicted in the other language. Experimental results on IWSLT and WMT datasets demonstrate that our method can obtain significant improvements over both conventional Neural Machine Translation (NMT) model and multilingual NMT model.

1 Introduction

Neural Machine Translation (NMT) based on the encoder-decoder framework has significantly improved translation quality due to its powerful end-to-end modeling (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017; Gehring et al., 2017; Hassan et al., 2018; Zhang and Zong, 2015). This paradigm facilitates the development of multilingual NMT (Dong et al., 2015; Luong et al., 2016; Johnson et al., 2017; Ha et al., 2016; Firat et al., 2016; Lakew et al., 2017; Tan et al., 2019; Wang et al., 2019), which handles multiple language pairs in one model, with the benefit of simplifying offline model training and easing online maintenance cost.

Although multilingual NMT attempts to utilize the complementary information of different languages (Lu et al., 2018; Neubig and Hu, 2018; Platanios et al., 2018; Wang et al., 2018), all of the models handle one language pair at each moment for both training and testing. However, we find that the generation process of different target languages can help each other. For example in Figure 1, when decoding the Chinese word "书" meaning "*book*" at step $t = 5$, the predicted Japanese word

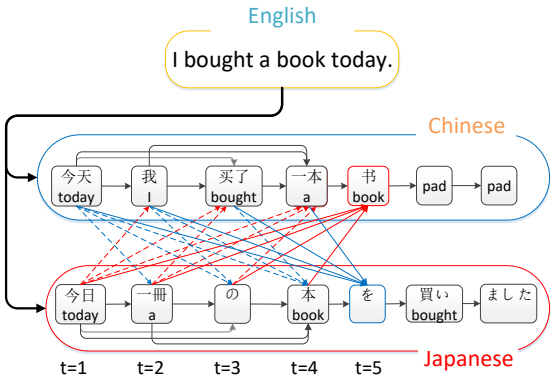

Figure 1: An example of an English sentence translated into Chinese and Japaneses sentences, in which two targets can interact with each other.

"本" with same meaning can provide the context at step $t = 4$. The reason is that the sentence structure between the two languages is different. It is *subject-verb-object* in Chinese while it is *subject-object-verb* in Japanese. Moreover, we find that two languages are complementary, and if decoders belonging to two different languages can interact with each other, the quality of translation will be improved.

In this work, we present a novel interactive decoding algorithm to generate two target languages simultaneously and interactively. To this end, we propose a synchronous attention model, in which the generation of one language can attend to already generated outputs of another language. As shown in Figure 1, the two decoders predict their outputs at the same time. At each moment, word prediction of each language does not only rely on previously generated targets itself but also depends on outputs of the other language.

We conduct extensive experiments to verify the effectiveness of our proposed approaches on English-to-German/French and English-to-Chinese/Japanese translation tasks.

*Jiajun Zhang is the corresponding author.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3350–3355, Hong Kong, China, November 3–7, 2019.

Our contributions in this work are two-fold:

(1) We propose a novel synchronous translation model that can predict outputs of two different languages simultaneously and interactively, which can enhance the translation quality of both languages.

(2) Extensive experiments show the superiority of our proposed method. Specifically, this synchronous approach can significantly outperform both the conventional NMT model and the multilingual NMT model.

2 Background

Owing to powerful modeling ability, our synchronous method relies on Transformer architecture (Vaswani et al., 2017), which is entirely based on the attention mechanism detailed below.

Scaled Dot-Product Attention: The inputs of attention mechanism contain queries Q, keys K, and values V. This function can be described as mapping a query and a set of key-value pairs to an output, and the output is calculated as a weighted sum of the values.

$$\text{Attention}\,(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

where d_k is dimension of keys, Q, K, V are obtained by linearly transforming input hidden states with projection matrices.

3 Synchronous Translation Method

As discussed in Sec. 1, outputs in different languages can be complementary and can help with each other. Thus, it is reasonable to improve translation quality with interactions of two decoders. In this section, we propose an interactive decoding algorithm and then describe how to implement it by a new attention mechanism, named as synchronous attention model.

3.1 Interactive Decoding Algorithm

Interactive decoding algorithm can generate translations of different languages in the same beam. At each step, each half of the beam produces translations in one target language conditioning on source sentence and the predicted tokens in both target languages. Here, we use two decoders with separated embeddings and softmax layers to operate two languages. The two decoders predict each token in parallel and keep interaction with each other, which can be formalized as follows:

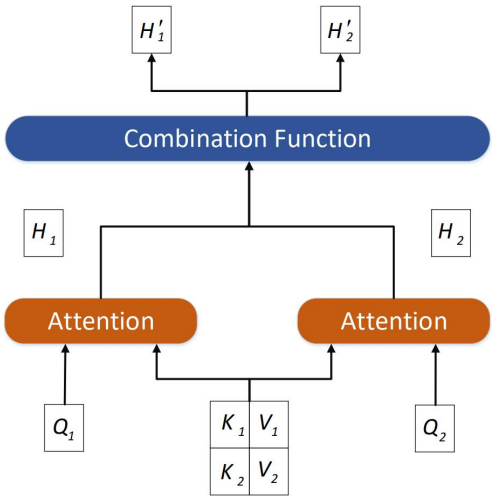

Figure 2: The synchronous self-attention framework, which simultaneously operates on the keys, queries and values of two different decoders.

$$\log P(\boldsymbol{y}^1|\boldsymbol{x}) = \log \prod_{i=0}^{n-1} p(y_i^1|\boldsymbol{x}, y_0^1, \ldots, y_{i-1}^1, y_0^2, \ldots, y_{i-1}^2)$$
$$\log P(\boldsymbol{y}^2|\boldsymbol{x}) = \log \prod_{i=0}^{n-1} p(y_i^2|\boldsymbol{x}, y_0^2, \ldots, y_{i-1}^2, y_0^1, \ldots, y_{i-1}^1) \quad (2)$$

where $\boldsymbol{x}$ is source sentence, $\boldsymbol{y}^1, \boldsymbol{y}^2$ are target sentences corresponding to two different languages. At time-step i, we have generated the first $i-1$ tokens of language-1 $\boldsymbol{y}^1$ and the first $i-1$ tokens of language-2 $\boldsymbol{y}^2$. Then both languages predictions can be utilized together with source sentence to generate tokens y_i^1 and y_i^2. This interaction between two languages is realized by synchronous attention model, which will be detailed in the following subsection.

It should be noted that the two language sentences can be generated in different directions (Liu et al., 2016; Zhang et al., 2018; Zhou et al., 2019), which means language-1 can be produced in left-to-right (L2R) manner while language-2 in right-to-left (R2L) manner. We will analyze the effect of different decoding manners in Sec. 5.2.

3.2 Synchronous Attention Model

Synchronous attention model (SyncAtt) is shown in Figure 2, in which inputs of two decoders contain queries (Q_1, Q_2), keys (K_1, K_2), and values (V_1, V_2) separately. The new hidden states (H_i') can be computed by our proposed synchronous atten-

tion as follows:

$$\begin{aligned} H_1^{'} &= \text{SyncAtt}\,(Q_1, [K_1; K_2], [V_1; V_2]) \\ H_2^{'} &= \text{SyncAtt}\,(Q_2, [K_1; K_2], [V_1; V_2]) \end{aligned} \tag{3}$$

where synchronous attention model (SyncAtt) can be described in detail:

$$\begin{aligned} H_1 &= \text{Attention}\,(Q_1, K_1, V_1) \\ \widetilde{H}_1 &= \text{Attention}\,(Q_1, K_2, V_2) \\ H_2 &= \text{Attention}\,(Q_2, K_2, V_2) \\ \widetilde{H}_2 &= \text{Attention}\,(Q_2, K_1, V_1) \\ H_1^{'} &= f(H_1; \widetilde{H}_1) = H_1 + \lambda \times \tanh(\widetilde{H}_1) \\ H_2^{'} &= f(H_2; \widetilde{H}_2) = H_2 + \lambda \times \tanh(\widetilde{H}_2) \end{aligned} \tag{4}$$

where λ is a balance weight of hidden states between two decoders decided by development set.

We apply our synchronous attention model to replace self-attention sub-layer in Transformer decoder, and it also utilizes the residual connections (He et al., 2016) around each sub-layer, followed by layer normalization (Ba et al., 2016).

3.3 Training

Since our synchronous method decodes two languages at the same time, the different decoders can be optimized simultaneously.

Supposing we have the trilingual datasets $\boldsymbol{D} = \{(\boldsymbol{x}, \boldsymbol{y}^1, \boldsymbol{y}^2)\}$, the objective function is to maximize the log-likelihood over the two target sentences:

$$L(\boldsymbol{\theta}) = \sum_{(\boldsymbol{x},\boldsymbol{y}^1,\boldsymbol{y}^2)\in D} \Big(\sum_{i=1}^{|y_i^1|} \log P(y_i^1|\boldsymbol{x}) + \sum_{i=1}^{|y_i^2|} \log P(y_i^2|\boldsymbol{x})\Big) \tag{5}$$

When calculating $P(y_i^1|\boldsymbol{x})$, except for the context from source side $\boldsymbol{x}$, our synchronous method employs not only previous reference $y_{<i}^1$ as condition, but the previous context of the other decoder reference $y_{<i}^2$. The calculation process of $P(y_i^2|\boldsymbol{x})$ is similar.

However, the practical situation is that the triple data is limited and hard to be collected. In this work, we construct the trilingual training corpus by data augmenting method (Sennrich et al., 2016a; Zhang and Zong, 2016). To achieve this, we first learn two independent translation models **Model-1** and **Model-2** on the bilingual training data $(\boldsymbol{x}^1, \boldsymbol{y}^1)$ and $(\boldsymbol{x}^2, \boldsymbol{y}^2)$ separately. Then, **Model-1** and **Model-2** are employed to decode the input sentences $\boldsymbol{x}^2$ and $\boldsymbol{x}^1$, resulting in pseudo training data $(\boldsymbol{x}^2, \boldsymbol{y}^{1\star})$ and $(\boldsymbol{x}^1, \boldsymbol{y}^{2\star})$, respectively. Thus, we can obtain the triple parallel training data $\boldsymbol{D} = \{(\boldsymbol{x}^1, \boldsymbol{y}^1, \boldsymbol{y}^{2\star})\} \cup \{(\boldsymbol{x}^2, \boldsymbol{y}^{1\star}, \boldsymbol{y}^2)\}$, which can be used to train our synchronous translation model mentioned above.

4 Experimental Settings

4.1 Data

We evaluate our proposed synchronous method on two translation tasks, which include English→Chinese/Japanese (briefly, En→Zh/Ja) and English→German/French (briefly, En→De/Fr) on IWSLT[1] datasets. The *IWSLT.TED.tst2013* and *IWSLT.TED.tst2014* are employed as devlopment set and test set respectively. Besides, we also perform En→De/Fr translation in large scale WMT14[2] datasets. We use *newstest2014* as test set.

En→Zh/Ja: For this translation task, the training sets of En→Zh and En→Ja consist of 231K, 223K sentence pairs. We tokenize the English sentences using a script from Moses (Koehn et al., 2007), and we segment Chinese and Japanese data by *jieba*[3] and *mecab*[4]. We use BPE method (Sennrich et al., 2016b) to encode the source side sentences and the combination of target side sentences respectively and limit the vocabularies of both sides to the most frequent 10k tokens.

En→De/Fr: We conduct this translation task on two different settings. One setting is using training set of IWSLT datasets which contains 206K sentence pairs for En→De and 233K sentence pairs for En→Fr. We follow the common practice to tokenize and lowercase all words. Sentences are encoded using BPE, which has a shared vocabulary of 10K tokens. At last, we construct pseudo triple data by the method described in Sec. 3.3. For the other setting, we extract the trilingual subset in WMT14 inspired by Zoph and Knight (2016), which includes about 2.43M sentence triples. We use 37K shared BPE tokens as vocabulary.

4.2 Training Details

We implement our synchronous translation based on the tensor2tensor[5] library. We train our models using the configuration *transformer_base* adopted

[1] https://wit3.fbk.eu
[2] http://www.statmt.org/wmt14/translation-task.html
[3] https://github.com/fxsjy/jieba
[4] http://taku910.github.io/mecab
[5] https://github.com/tensorflow/tensor2tensor

Hyper-paramter (λ)	En-De/Fr		En-Zh/Ja	
	En-De	En-Fr	En-Zh	En-Ja
0.1	**30.95**	**43.01**	**16.33**	**18.88**
0.2	30.77	42.99	16.06	18.82
0.3	30.55	42.99	15.96	18.38
0.4	29.60	42.52	15.66	18.03
0.5	29.19	41.87	15.17	17.42

Table 1: Experiment results on the development set with different λs.

Method	En-Zh/Ja		En-De/Fr	
	En-Zh	En-Ja	En-De	En-Fr
Indiv	15.68	16.56	27.11	40.62
Indiv + pseudo	16.72	18.02	28.47	40.39
Multi	17.06	18.31	27.79	40.97
Multi + pseudo	17.10	18.40	28.56	40.62
SyncTrans	**17.97**	**19.31**	**29.16**	**41.53**

Table 2: Translation performance on IWSLT datasets. *SyncTrans* represents our proposed synchronous translation method. All results of our *SyncTrans* are significantly better than both *Indiv* and *Multi* ($p < 0.01$).

by Vaswani et al. (2017), which contains a 6-layer encoder and a 6-layer decoder with 512-dimensional hidden representations. During training, each mini-batch contains roughly 4,096 tokens for both source and target sides. We use Adam optimizer (Kingma and Ba, 2014) with β_1=0.9, β_2=0.98, and ϵ=10^{-9}. For decoding, we set beam size to be $k = 4$ and length penalty $\alpha = 0.6$. All our methods are trained and tested on single Nvidia P40 GPU.

We investigate the impact of different λs in our synchronous attention model. As shown in Table 1, when λ=0.1, the translation results perform best on development set for both En→Zh/Ja and En→De/Fr tasks, and we will use this setting in the subsequent experiments.

5 Results and Analysis

The translation performance of IWSLT datasets is evaluated by case-insensitive BLEU4 (Papineni et al., 2002) for En→De/Fr task and character-level BLEU5 for En→Zh/Ja task. For WMT14 datasets, we calculate the case-sensitive BLEU4 the same as previous work. In our experiments, the NMT models trained on individual language pair are denoted by *Indiv*.

5.1 Results on IWSLT

Table 2 shows the main translation results of En→Zh/Ja and En→De/Fr on IWSLT datasets. We also conduct a typical one-to-many translation adopting Johnson et al. (2017) method on Transformer as our another baseline model, referred to *Multi*. Compared with *Indiv*, we can see that *Multi* achieves better results on all cases, which can be attributed to that the encoder can be enhanced by extra training data from the other language pair.

As for our proposed method, the synchronous translation method performs significantly better than both *Indiv* and *Multi* baseline methods, and it can achieve the improvements up to 2.75 BLEU points (19.31 vs. 16.56) on En→Ja.

To perform synchronous translation, the triple parallel corpus contains pseudo training data we construct. For a fair comparison, we also conduct our baseline methods on the training sentences augmented by the pseudo corpus. From row 2 and row 4 in Table 2, our method achieves better performance than both *Multi + pseudo* method and *Indiv + pseudo* method with gains of 0.82 BLEU and 1.09 points on average, which demonstrates the effectiveness of our method.

5.2 L2R or R2L Manner

As described in Sec. 3.1, two target languages can be generated in L2R or R2L manner, which can provide the future contexts for each other. We further conduct an experiment to investigate different decoding manners in this work.

Figure 3 reports the results. We observe that when performing En→De/Fr translation, one language generated from R2L manner is helpful for the other language but do harm to itself. However, for En→Zh/Ja translation, Japanese can achieve improvements on both L2R and R2L decoding settings. The reason is that Japanese is suited for translating from right-to-left manner, and it can take advantage of future outputs from Chinese.

5.3 Results on WMT

We also employ our method on the real triple training datasets, which can be collected from WMT14 En→De/Fr datasets as described in Sec. 4.1. From Table 3, we observe that our method consistently outperforms baseline models. Note that in contrast to results on IWSLT datasets, *Multi* can not perform on par with *Indiv*, because the source side data for two language pairs are the same, and encoder network can not be enhanced as *Multi* method in Sec. 5.1.

Moreover, we construct a large scale pseudo

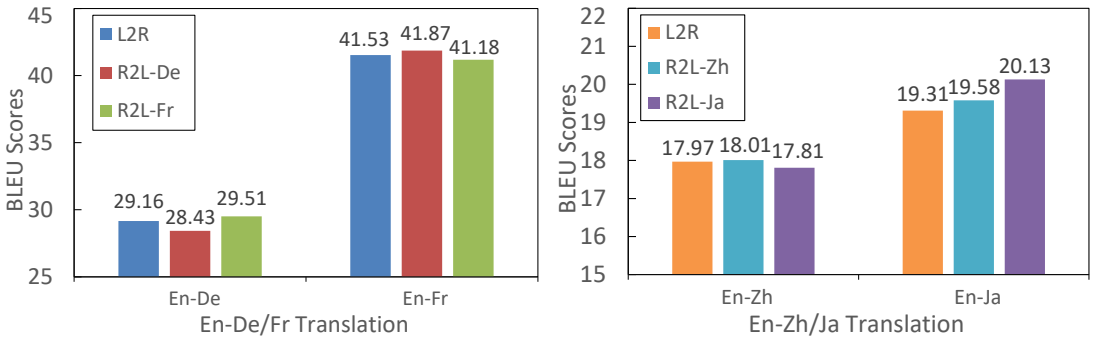

Figure 3: The comparison of L2R and R2L decoding manners, in which one of target languages generated from R2L manner.

Method	WMT14 (2.43M)		WMT14 (4.50M)
	En-De	En-Fr	En-De
Indiv	24.33	37.12	26.53
Multi	23.46	36.33	25.81
SyncTrans	**24.84**$^{\dagger *}$	**37.66**$^{\dagger *}$	**27.01**$^{\dagger *}$

Table 3: Translation quality of En-De/Fr on WMT14 datasets. The significance values with respect to the baseline method *Indiv* and *Multi* method are denoted by "$*$" and "†" respectively, indicating our proposed *SyncTrans* significantly better than both *Indiv* ($p < 0.05$) and *Multi* ($p < 0.01$) methods.

triple data about 4.5M[6]. The result is demonstrated in the last column of Table 3, in which our synchronous method performs better than the baseline methods as well.

6 Conclusion

In this paper, we propose an interactive decoding algorithm to generate two target languages simultaneously and interactively. The empirical experiments on four language pairs demonstrate that our approach can obtain significant improvements over both the NMT model trained on individual language pair and multilingual NMT model. For the future work, we plan to extend our method on more than two target languages and explore other effective interactive approaches to improve the translation quality further.

Acknowledgments

The research work described in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2016QY02D0303, the Natural Science Foundation of China under Grant No. U1836221 and the Beijing Municipal Science and Technology Project under Grant No. Z181100008918017. The research work in this paper has also been supported by Beijing Advanced Innovation Center for Language Resources. We thank the three anonymous reviewers for their efforts on this manuscript. We would like to thank Yanfen Zhang and Junnan Zhu for their invaluable discussions on this paper.

[6]The En-Fr part is entirely created by translating English of WMT14 En-De datasets into French using the model trained on WMT14 En-Fr corpus about 37M.

References

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*.

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*.

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

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *In proceddings of CVPR 2016*.

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.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. *In Proceedings of ACL 2007*.

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

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Agreement on target-bidirectional neural machine translation. *In Proceedings of NAACL 2016*, pages 411–416.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. *arXiv preprint arXiv:1804.08198*.

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*.

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.

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.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *In Proceedings of ICLR 2019*.

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.

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019. A compact and language-sensitive multilingual translation method. *In Proceedings of ACL 2019*, pages 1213–1223.

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.

Xiangwen Zhang, Jinsong Su, Yue Qin, Yang Liu, Rongrong Ji, and Hongji Wang. 2018. Asynchronous bidirectional decoding for neural machine translation. *In proceddings of AAAI 2018*.

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.