Regularized Context Gates on Transformer for Machine Translation

Xintong Li[1], Lemao Liu[2], Rui Wang[3], Guoping Huang[2], Max Meng[1]

[1]The Chinese University of Hong Kong   [2]Tencent AI Lab
[3]National Institute of Information and Communications Technology
znculee@gmail.com   redmondliu@tencent.com   wangrui@nict.go.jp
donkeyhuang@tencent.com   max.meng@cuhk.edu.hk

Abstract

Context gates are effective to control the contributions from the source and target contexts in the recurrent neural network (RNN) based neural machine translation (NMT). However, it is challenging to extend them into the advanced Transformer architecture, which is more complicated than RNN. This paper first provides a method to identify source and target contexts and then introduce a gate mechanism to control the source and target contributions in Transformer. In addition, to further reduce the bias problem in the gate mechanism, this paper proposes a regularization method to guide the learning of the gates with supervision automatically generated using pointwise mutual information. Extensive experiments on 4 translation datasets demonstrate that the proposed model obtains an averaged gain of 1.0 BLEU score over a strong Transformer baseline.

1 Introduction

An essence to modeling translation is how to learn an effective context from a sentence pair. Statistical machine translation (SMT) models the source context from the source-side of a translation model and models the target context from a target-side language model (Koehn et al., 2003; Koehn, 2009; Chiang, 2005). These two models are trained independently. On the contrary, neural machine translation (NMT) advocates a unified manner to jointly learn source and target context using an encoder-decoder framework with an attention mechanism, leading to substantial gains over SMT in translation quality (Sutskever et al., 2014; Bahdanau et al., 2014; Gehring et al., 2017; Vaswani et al., 2017). Prior work on attention mechanism (Luong et al., 2015; Liu et al., 2016; Mi et al., 2016; Chen et al., 2018; Li et al., 2018; Elbayad et al., 2018; Yang et al., 2020) have shown a better context representation is helpful to translation performance.

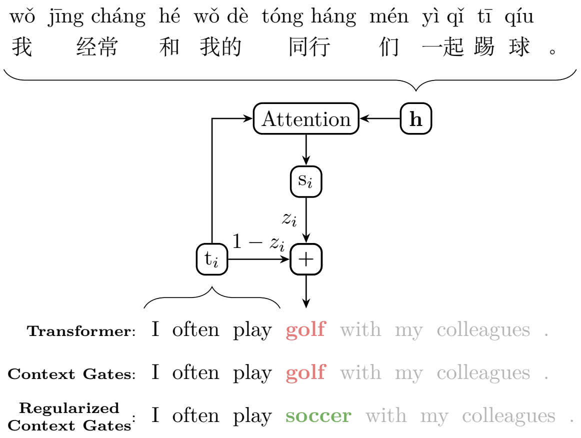
wǒ jīng cháng hé wǒ dè tóng háng mén yì qǐ tī qíu
我 经常 和 我的 同行 们 一起 踢 球 。
Attention
h
s_i
z_i
$1 - z_i$
t_i
+
Transformer: I often play golf with my colleagues .
Context Gates: I often play golf with my colleagues .
Regularized Context Gates: I often play soccer with my colleagues .

Figure 1: A running example to raise the context control problem. Both original and context gated Transformer obtain an unfaithful translation by wrongly translate "*tī qíu*" into "*play golf*" because referring too much target context. By regularizing the context gates, the purposed method corrects the translation of "*tī qíu*" into "*play soccer*". The light font denotes the target words to be translated in the future. For original Transformer, the source and target context are added directly without any rebalancing.

However, a standard NMT system is incapable of effectively controlling the contributions from source and target contexts (He et al., 2018) to deliver highly adequate translations as shown in Figure 1. As a result, Tu et al. (2017) carefully designed context gates to dynamically control the influence from source and target contexts and observed significant improvements in the recurrent neural network (RNN) based NMT. Although Transformer (Vaswani et al., 2017) delivers significant gains over RNN for translation, there are still one third translation errors related to context control problem as described in Section 3.3. Obviously, it is feasible to extend the context gates in RNN based NMT into Transformer, but an obstacle to accomplishing this goal is the complicated archi-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8555–8562
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

tecture in Transformer, where the source and target words are tightly coupled. Thus, it is challenging to put context gates into practice in Transformer.

In this paper, under the Transformer architecture, we firstly provide a way to define the source and target contexts and then obtain our model by combining both source and target contexts with context gates, which actually induces a probabilistic model indicating whether the next generated word is contributed from the source or target sentence (Li et al., 2019). In our preliminary experiments, this model only achieves modest gains over Transformer because the context selection error reduction is very limited as described in Section 3.3. To further address this issue, we propose a probabilistic model whose loss function is derived from external supervision as regularization for the context gates. This probabilistic model is jointly trained with the context gates in NMT. As it is too costly to annotate this supervision for a large-scale training corpus manually, we instead propose a simple yet effective method to automatically generate supervision using pointwise mutual information, inspired by word collocation (Bouma, 2009). In this way, the resulting NMT model is capable of controlling the contributions from source and target contexts effectively.

We conduct extensive experiments on 4 benchmark datasets, and experimental results demonstrate that the proposed gated model obtains an averaged improvement of 1.0 BLEU point over corresponding strong Transformer baselines. In addition, we design a novel analysis to show that the improvement of translation performance is indeed caused by relieving the problem of wrongly focusing on the source or target context.

2 Methodology

Given a source sentence $\mathbf{x} = \langle \mathrm{x}_1, \cdots, \mathrm{x}_{|\mathbf{x}|} \rangle$ and a target sentence $\mathbf{y} = \langle \mathrm{y}_1, \cdots, \mathrm{y}_{|\mathbf{y}|} \rangle$, our proposed model is defined by the following conditional probability under the Transformer architecture: [1]

$$P(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{|\mathbf{y}|} P(\mathrm{y}_i \mid \mathbf{y}_{<i}, \mathbf{x}) = \prod_{i=1}^{|\mathbf{y}|} P\left(\mathrm{y}_i \mid \mathrm{c}_i^L\right), \tag{1}$$

where $\mathbf{y}_{<i} = \langle \mathrm{y}_1, \ldots, \mathrm{y}_{i-1} \rangle$ denotes a prefix of $\mathbf{y}$ with length $i-1$, and c_i^L denotes the L^{th} layer context in the decoder with L layers which is obtained from the representation of $\mathbf{y}_{<i}$ and $\mathbf{h}^L$, i.e., the top layer hidden representation of $\mathbf{x}$, similar to the original Transformer. To finish the overall definition of our model in equation 1, we will expand the definition c_i^L based on context gates in the following subsections.

2.1 Context Gated Transformer

To develop context gates for our model, it is necessary to define the source and target contexts at first. Unlike the case in RNN, the source sentence $\mathbf{x}$ and the target prefix $\mathbf{y}_{<i}$ are tightly coupled in our model, and thus it is not trivial to define the source and target contexts.

Suppose the source and target contexts at each layer l are denoted by s_i^l and t_i^l. We recursively define them from $\mathbf{c}_{<i}^{l-1}$ as follows. [2]

$$\begin{aligned} \mathrm{t}_i^l &= \mathrm{rn} \circ \mathrm{ln} \circ \mathrm{att}\left(\mathrm{c}_i^{l-1}, \mathbf{c}_{<i}^{l-1}\right), \\ \mathrm{s}_i^l &= \mathrm{ln} \circ \mathrm{att}\left(\mathrm{t}_i^l, \mathbf{h}^L\right), \end{aligned} \tag{2}$$

where $\circ$ is functional composition, $\mathrm{att}\,(\mathrm{q}, \mathrm{kv})$ denotes multiple head attention with q as query, k as key, v as value, and rn as a residual network (He et al., 2016), ln is layer normalization (Ba et al., 2016), and all parameters are removed for simplicity.

In order to control the contributions from source or target side, we define c_i^l by introducing a context gate $\boldsymbol{z}_i^l$ to combine s_i^l and t_i^l as following:

$$\mathbf{c}_i^l = \mathrm{rn} \circ \mathrm{ln} \circ \mathrm{ff}\left((\mathbf{1} - \boldsymbol{z}_i^l) \otimes \mathrm{t}_i^l + \boldsymbol{z}_i^l \otimes \mathrm{s}_i^l\right) \tag{3}$$

with

$$\boldsymbol{z}_i^l = \sigma\left(\mathrm{ff}\left(\mathrm{t}_i^l \| \mathrm{s}_i^l\right)\right), \tag{4}$$

where ff denotes a feedforward neural network, $\|$ denotes concatenation, $\sigma(\cdot)$ denotes a sigmoid function, and $\otimes$ denotes an element-wise multiplication. $\boldsymbol{z}_i^l$ is a vector (Tu et al. (2017) reported that a gating vector is better than a gating scalar). Note that each component in $\boldsymbol{z}_i^l$ actually induces a probabilistic model indicating whether the next generated word y_i is mainly contributed from the source ($\mathbf{x}$) or target sentence ($\mathbf{y}_{<i}$) , as shown in Figure 1.

Remark It is worth mentioning that our proposed model is similar to the standard Transformer with boiling down to replacing a residual connection

[1] Throughout this paper, a variable in bold font such as $\mathbf{x}$ denotes a sequence while regular font such as x denotes an element which may be a scalar x, vector $\boldsymbol{x}$ or matrix $\boldsymbol{X}$.

[2] For the base case, $\mathbf{c}_{<i}^0$ is word embedding of $\mathbf{y}_{<i}$.

with a high way connection (Srivastava et al., 2015; Zhang et al., 2018): if we replace $(\mathbf{1}-\boldsymbol{z}_i^l)\otimes \mathrm{t}_i^l + \boldsymbol{z}_i^l \otimes \mathrm{s}_i^l$ in equation 3 by $\mathbf{t}_i^l + \mathbf{s}_i^l$, the proposed model is reduced to Transformer.

2.2 Regularization of Context Gates

In our preliminary experiments, we found learning context gates from scratch cannot effectively reduce the context selection errors as described in Section 3.3.

To address this issue, we propose a regularization method to guide the learning of context gates by external supervision z_i^* which is a binary number representing whether y_i is contributed from either source ($z_i^* = 1$) or target sentence ($z_i^* = 0$). Formally, the training objective is defined as follows:

$$\ell = -\log P(\mathbf{y} \mid \mathbf{x}) + \lambda \sum_{l,i} \Big(z_i^* \max(\mathbf{0.5} - \boldsymbol{z}_i^l, \mathbf{0}) + (1 - z_i^*) \max(\boldsymbol{z}_i^l - \mathbf{0.5}, \mathbf{0}) \Big), \quad (5)$$

where $\boldsymbol{z}_i^l$ is a context gate defined in equation 4 and λ is a hyperparameter to be tuned in experiments. Note that we only regularize the gates during the training, but we skip the regularization during inference.

Because golden z_i^* are inaccessible for each word y_i in the training corpus, we ideally have to annotate it manually. However, it is costly for human to label such a large scale dataset. Instead, we propose an automatic method to generate its value in practice in the next subsection.

2.3 Generating Supervision z_i^*

To decide whether y_i is contributed from the source ($\mathbf{x}$) or target sentence ($\mathbf{y}_{<i}$) (Li et al., 2019), a metric to measure the correlation between a pair of words ($\langle \mathrm{y}_i, \mathrm{x}_j \rangle$ or $\langle \mathrm{y}_i, \mathrm{y}_k \rangle$ for $k < i$) is first required. This is closely related to a well-studied problem, i.e., word collocation (Liu et al., 2009), and we simply employ the pointwise mutual information (PMI) to measure the correlation between a word pair $\langle \mu, \nu \rangle$ following Bouma (2009):

$$\begin{aligned} \mathrm{pmi}\left(\mu, \nu\right) &= \log \frac{P(\mu,\nu)}{P(\mu)P(\nu)} \\ &= \log Z + \log \frac{C(\mu,\nu)}{C(\mu)C(\nu)}, \end{aligned} \quad (6)$$

where $C\left(\mu\right)$ and $C\left(\nu\right)$ are word counts, $C\left(\mu,\nu\right)$ is the co-occurrence count of words μ and ν, and Z is the normalizer, i.e., the total number of all possible (μ, ν) pairs. To obtain the context gates, we define two types of PMI according to different $C\left(\mu,\nu\right)$ including two scenarios as follows.

PMI in the Bilingual Scenario For each parallel sentence pair $\langle \mathbf{x}, \mathbf{y} \rangle$ in training set, $C\left(\mathrm{y}_i, \mathrm{x}_j\right)$ is added by one if both $\mathrm{y}_i \in \mathbf{y}$ and $\mathrm{x}_j \in \mathbf{x}$.

PMI in the Monolingual Scenario In the translation scenario, only the words in the preceding context of a target word should be considered. So for any target sentence $\mathbf{y}$ in the training set, $C\left(\mathrm{y}_i, \mathrm{y}_k\right)$ is added by one if both $\mathrm{y}_i \in \mathbf{y}$ and $\mathrm{y}_k \in \mathbf{y}_{<i}$.

Given the two kinds of PMI for a bilingual sentence $\langle \mathbf{x}, \mathbf{y} \rangle$, each z_i^* for each y_i is defined as follows,

$$z_i^* = \mathbb{1}_{\max_j \mathrm{pmi}(\mathrm{y}_i, \mathrm{x}_j) > \max_{k<i} \mathrm{pmi}(\mathrm{y}_i, \mathrm{y}_k)}, \quad (7)$$

where $\mathbb{1}_b$ is a binary function valued by 1 if b is true and 0 otherwise. In equation 7, we employ $\max$ strategy to measure the correlation between y_i and a sentence ($\mathbf{x}$ or $\mathbf{y}_{<i}$). Indeed, it is similar to use the average strategy, but we did not find its gains over $\max$ in our experiments.

3 Experiments

The proposed methods are evaluated on NIST ZH⇒EN [3], WMT14 EN⇒DE [4], IWSLT14 DE⇒EN [5] and IWSLT17 FR⇒EN [6] tasks. To make our NMT models capable of open-vocabulary translation, all datasets are preprocessed with Byte Pair Encoding (Sennrich et al., 2015). All proposed methods are implemented on top of Transformer (Vaswani et al., 2017) which is the state-of-the-art NMT system. Case-insensitive BLEU score (Papineni et al., 2002) is used to evaluate translation quality of ZH⇒EN, DE⇒EN and FR⇒EN. For the fair comparison with the related work, EN⇒DE is evaluated with case-sensitive BLEU score. Setup details are described in Appendix A.

3.1 Tuning Regularization Coefficient

In the beginning of our experiments, we tune the regularization coefficient λ on the DE⇒EN task. Table 2 shows the robustness of λ, because the translation performance only fluctuates slightly over various λ. In particular, the best performance

[3] LDC2000T50, LDC2002L27, LDC2002T01, LDC2002E18, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07

[4] WMT14: http://www.statmt.org/wmt14/

[5] IWSLT14: http://workshop2014.iwslt.org/

[6] IWSLT17: http://workshop2017.iwslt.org/

Models		params $\times 10^6$	ZH⇒EN			EN⇒DE	DE⇒EN	FR⇒EN
			MT05	MT06	MT08			
RNN based NMT		84	30.6	31.1	23.2	–	–	–
Tu et al. (2017)		88	34.1	34.8	26.2	–	–	–
Vaswani et al. (2017)		65	–	–	–	27.3	–	–
Ma et al. (2018)		–	36.8	35.9	27.6	–	–	–
Zhao et al. (2018)		–	43.9	44.0	33.3	–	–	–
Cheng et al. (2018)		–	44.0	44.4	34.9	–	–	–
Transformer		74	46.9	47.4	38.3	27.4	32.2	36.8
This Work	Context Gates	92	47.1	47.6	39.1	27.9	32.5	37.7
	Regularized Context Gates	92	**47.7**	**48.3**	**39.7**	**28.1**	**33.0**	**38.3**

Table 1: Translation performances (BLEU). The RNN based NMT (Bahdanau et al., 2014) is reported from the baseline model in Tu et al. (2017). "params" shows the number of parameters of models when training ZH⇒EN except Vaswani et al. (2017) is for EN⇒DE tasks.

λ	0.1	0.5	1	2	10
BLEU	32.7	32.6	**33.0**	32.7	32.6

* Results are measured on DE⇒EN task.

Table 2: Translation performance over different regularization coefficient λ.

is achieved when $\lambda = 1$, which is the default setting throughout this paper.

3.2 Translation Performance

Table 1 shows the translation quality of our methods in BLEU. Our observations are as follows:

1) The performance of our implementation of the Transformer is slightly higher than Vaswani et al. (2017), which indicates we are in a fair comparison.

2) The proposed Context Gates achieves modest improvement over the baseline. As we mentioned in Section 2.1, the structure of RNN based NMT is quite different from the Transformer. Therefore, naively introducing the gate mechanism to the Transformer without adaptation does not obtain similar gains as it does in RNN based NMT.

3) The proposed Regularized Context Gates improves nearly 1.0 BLEU score over the baseline and outperforms all existing related work. This indicates that the regularization can make context gates more effective in relieving the context control problem as discussed following.

3.3 Error Analysis

To explain the success of Regularized Context Gates, we analyze the error rates of translation and context selection. Given a sentence pair $\mathbf{x}$ and $\mathbf{y}$, the forced decoding translation error is defined as $P(\mathrm{y}_i \mid \mathbf{y}_{<i}, \mathbf{x}) < P(\hat{\mathrm{y}}_i \mid \mathbf{y}_{<i}, \mathbf{x})$, where $\hat{\mathrm{y}}_i \triangleq \arg\max_{\mathrm{v}} P(\mathrm{v} \mid \mathbf{y}_{<i}, \mathbf{x})$ and v denotes any token in the vocabulary. The context selection error is defined as $z_i^*(\mathrm{y}_i) \neq z_i^*(\hat{\mathrm{y}}_i)$, where z_i^* is defined in equation 7. Note that a context selection error must be a translation error but the opposite is not true. The example shown in Figure 1 also demonstrates a context selection error indicating the translation error is related with the bad context selection.

Models	FER	CER	CE/FE
Transformer	40.5	13.8	33.9
Context Gates	40.5	13.7	33.7
Regularized Context Gates	**40.0**	**13.4**	**33.4**

* Results are measured on MT08 of ZH⇒EN task.

Table 3: Forced decoding translation error rate (**FER**), context selection error rate (**CER**) and the proportion of context selection errors over forced decoding translation errors (**CE/FE**) of the original and context gated Transformer with or without regularization.

As shown in Table 3, the Regularized Context Gates significantly reduce the translation error by avoiding the context selection error. The Context Gates are also able to avoid few context selection error but cannot make a notable improvement in translation performance. It is worth to note that there is approximately one third translation error is related to context selection error. The Regularized Context Gates indeed alleviate this severe problem by effectively rebalancing of source and target context for translation.

3.4 Statistics of Context Gates

Table 4 summarizes the mean and variance of each context gate (every dimension of the context gate vectors) over the MT08 test set. It shows that learning context gates freely from scratch tends to pay more attention to target context (0.38 < 0.5), which

Models	Mean	Variance
Context Gates	0.38	0.10
Regularized Context Gates	0.51	0.13

* Results are measured on MT08 of ZH⇒EN task.

Table 4: Mean and variance of context gates

means the model tends to trust its language model more than the source context, and we call this context imbalance bias of the freely learned context gate. Specifically, this bias will make the translation unfaithful for some source tokens. As shown in Table 4, the Regularized Context Gates demonstrates more balanced behavior (0.51≈0.5) over the source and target context with similar variance.

3.5 Regularization in Different Layers

To investigate the sensitivity of choosing different layers for regularization, we only regularize the context gate in every single layer. Table 5 shows that there is no significant performance difference, but all single layer regularized context gate models are slightly inferior to the model, which regularizes all the gates. Moreover, since nearly no computation overhead is introduced and for design simplicity, we adopt regularizing all the layers.

Layers	N/A	1	2	3	4	ALL
BLEU	32.5	32.8	32.7	32.5	32.3	33.0

* Results are measured on DE⇒EN task.

Table 5: Regularize context gates on different layers."N/A" indicates regularization is not added. "ALL" indicates regularization is added to all the layers.

3.6 Effects on Long Sentences

In Tu et al. (2017), context gates alleviate the problem of long sentence translation of attentional RNN based system (Bahdanau et al., 2014). We follow Tu et al. (2017) and compare the translation performances according to different lengths of the sentences. As shown in Figure 2, we find Context Gates does not improve the translation of long sentences but translate short sentences better. Fortunately, the Regularized Context Gates indeed significantly improves the translation for both short sentences and long sentences.

4 Conclusions

This paper transplants context gates from the RNN based NMT to the Transformer to control the source and target context for translation. We find that context gates only modestly improve the translation quality of the Transformer, because learning context gates freely from scratch is more challenging for the Transformer with the complicated structure than for RNN. Based on this observation, we propose a regularization method to guide the learning of context gates with an effective way to generate supervision from training data. Experimental results show the regularized context gates can significantly improve translation performances over different translation tasks even though the context control problem is only slightly relieved. In the future, we believe more work on alleviating context control problem has the potential to improve translation performance as quantified in Table 3.

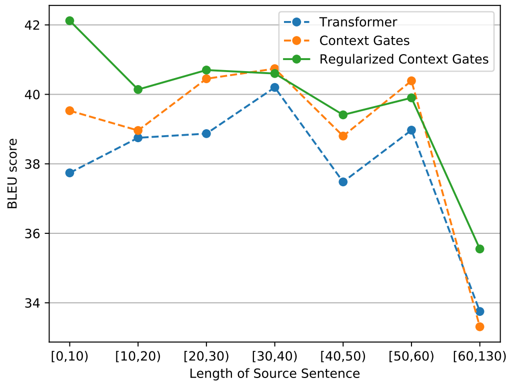

Figure 2: Translation performance on MT08 test set with respect to different lengths of source sentence. Regularized Context Gates significantly improves the translation of short and long sentences.

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. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2018. Syntax-directed attention for neural machine translation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. *arXiv preprint arXiv:1805.06130*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2018. Pervasive attention: 2d convolutional neural networks for sequence-to-sequence prediction. *arXiv preprint arXiv:1808.03867*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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 *Advances in Neural Information Processing Systems*, pages 7944–7954.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1293–1303.

Xintong Li, Lemao Liu, Zhaopeng Tu, Shuming Shi, and Max Meng. 2018. Target foresight based attention for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1380–1390.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. *arXiv preprint arXiv:1609.04186*.

Zhanyi Liu, Haifeng Wang, Hua Wu, and Sheng Li. 2009. Collocation extraction using monolingual word alignment method. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 487–495. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Shuming Ma, Xu Sun, Yizhong Wang, and Junyang Lin. 2018. Bag-of-words as target for neural machine translation. *arXiv preprint arXiv:1805.04871*.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised attentions for neural machine translation. *arXiv preprint arXiv:1608.00112*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Mingming Yang, Min Zhang, Kehai Chen, Rui Wang, and Tiejun Zhao. 2020. Neural machine translation with target-attention model. *IEICE TRANSACTIONS on Information and Systems*, 103(3):684–694.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, and Yang Liu. 2017. Thumt: an open source toolkit for neural machine translation. *arXiv preprint arXiv:1706.06415*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. *arXiv preprint arXiv:1810.03581*.

Yang Zhao, Jiajun Zhang, Zhongjun He, Chengqing Zong, and Hua Wu. 2018. Addressing troublesome words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 391–400.

A Details of Data and Implementation

The training data for ZH⇒EN task consists of 1.8M sentence pairs. The development set is chosen as NIST02 and test sets are NIST05, 06, 08. For EN⇒DE task, its training data contains 4.6M sentences pairs. Both FR⇒EN and DE⇒EN tasks contain around 0.2M sentence pairs. For ZH⇒EN and EN⇒DE tasks, the joint vocabulary is built with 32K BPE merge operations, and for DE⇒EN and FR⇒EN tasks it is built with 16K merge operations.

Our implementation of context gates and the regularization are based on Transformer, implemented by THUMT (Zhang et al., 2017). For ZH⇒EN and EN⇒DE tasks, only the sentences of length up to 256 tokens are used with no more than 2^{15} tokens in a batch. The dimension of both word embeddings and hidden size are 512. Both encoder and decoder have 6 layers and adopt multi-head attention with 8 heads. For FR⇒EN and DE⇒EN tasks, we use a smaller model with 4 layers and 4 heads, and both the embedding size and the hidden size is 256. The training batch contains no more than 2^{12} tokens. For all tasks, the beam size for decoding is 4, and the loss function is optimized with Adam, where $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$.