Regularizing Neural Machine Translation by Target-Bidirectional Agreement

Zhirui Zhang†*, Shuangzhi Wu‡, Shujie Liu§, Mu Li§, Ming Zhou§, Tong Xu††

††University of Science and Technology of China, Hefei, China
‡‡Harbin Institute of Technology, Harbin, China
§§Microsoft Research Asia
†zrustc11@gmail.com ‡tongxu@ustc.edu.cn
§§{v-shuawu,shujliu,muli,mingzhou}@microsoft.com

Abstract

Although Neural Machine Translation (NMT) has achieved remarkable progress in the past several years, most NMT systems still suffer from a fundamental shortcoming as in other sequence generation tasks: errors made early in generation process are fed as inputs to the model and can be quickly amplified, harming subsequent sequence generation. To address this issue, we propose a novel model regularization method for NMT training, which aims to improve the agreement between translations generated by left-to-right (L2R) and right-to-left (R2L) NMT decoders. This goal is achieved by introducing two Kullback-Leibler divergence regularization terms into the NMT training objective to reduce the mismatch between output probabilities of L2R and R2L models. In addition, we also employ a joint training strategy to allow L2R and R2L models to improve each other in an interactive update process. Experimental results show that our proposed method significantly outperforms state-of-the-art baselines on Chinese-English and English-German translation tasks.

Introduction

Neural Machine Translation (NMT) (Cho et al. 2014; Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2014) has seen the rapid development in the past several years, from catching up with Statistical Machine Translation (SMT) (Koehn, Och, and Marcu 2003; Chiang 2007) to outperforming it by significant margins on many languages (Sennrich, Haddow, and Birch 2016; Wu et al. 2016; Tu et al. 2016; Eriguchi, Hashimoto, and Tsuruoka 2016; Wang et al. 2017a; Vaswani et al. 2017). In a conventional NMT model, an encoder first transforms the source sequence into a sequence of intermediate hidden vector representations, based on which, a decoder generates the target sequence word by word.

Due to the autoregressive structure, current NMT systems usually suffer from the so-called exposure bias problem (Bengio et al. 2015): during inference, true previous target tokens are unavailable and replaced by tokens generated by the model itself, thus mistakes made early can mislead subsequent translation, yielding unsatisfied translations with good prefixes but bad suffixes (shown in Table 1). Such an issue can become severe as sequence length increases.

To address this problem, one line of research attempts to reduce the inconsistency between training and inference so as to improve the robustness when giving incorrect previous predictions, such as designing sequence-level objectives or adopting reinforcement learning approaches (Ranzato et al. 2015; Shen et al. 2016; Wiseman and Rush 2016). Another line tries to leverage a complementary NMT model that generates target words from right to left (R2L) to distinguish unsatisfied translation results from a n-best list generated by the L2R model (Liu et al. 2016; Wang et al. 2017b).

In their work, the R2L NMT model is only used to re-rank the translation candidates generated by L2R model, while the candidates in the n-best list still suffer from the exposure bias problem and limit the room for improvement. Another problem is that, the complementary R2L model tends to generate translation results with good suffixes and bad prefixes, due to the same exposure bias problem, as shown in Table 1. Similar with using the R2L model to augment the L2R model, the L2R model can also be leveraged to improve the R2L model.

Instead of re-ranking the n-best list, we try to take consideration of the agreement between the L2R and R2L models into both of their training objectives, hoping that the agreement information can help to learn better models integrating their advantages to generate translations with good prefixes.
and good suffixes. To this end, we introduce two Kullbach-Leibler (KL) divergences between the probability distributions defined by L2R and R2L models into the NMT training objective as regularization terms. Thus, we can not only maximize the likelihood of training data but also minimize the L2R and R2L model divergence at the same time, in which the latter one serves as a measure of exposure bias problem of the currently evaluated model. With this method, the L2R model can be enhanced using the R2L model as a helper system, and the R2L model can also be improved with the help of L2R model. We integrate the optimization of R2L and L2R models into a joint training framework, in which they act as helper systems for each other, and both models achieve further improvements with an interactive update process.

Our experiments are conducted on Chinese-English and English-German translation tasks, and demonstrate that our proposed method significantly outperforms state-of-the-art baselines.

### Neural Machine Translation

Neural Machine Translation (NMT) is an end-to-end framework to directly model the conditional probability \( P(y|x) \) of target translation \( y = (y_1, y_2, \ldots, y_T) \) given source sentence \( x = (x_1, x_2, \ldots, x_T) \). In practice, NMT systems are usually implemented with an attention-based encoder-decoder architecture. The encoder reads the source sentence \( x \) and transforms it into a sequence of intermediate hidden vectors \( h = (h_1, h_2, \ldots, h_T) \) using a neural network. Given the hidden state \( h \), the decoder generates target translation \( y \) with another neural network that jointly learns language and alignment models.

The structure of neural networks first employs recurrent neural networks (RNN) or its variants - Gated Recurrent Unit (Cho et al. 2014) and Long Short-Term Memory (Hochreiter and Schmidhuber 1997) networks. Recently, two additional architectures have been proposed, improving not only parallelization but also the state-of-the-art result: the fully convolutional model (Gehring et al. 2017) and the self-attentional transformer (Vaswani et al. 2017).

For model training, given a parallel corpus \( D = \{(x^{(n)}, y^{(n)})\}_{n=1}^N \), the standard training objective in NMT is to maximize the likelihood of the training data:

\[
L(\theta) = \sum_{n=1}^N \log P(y^{(n)}|x^{(n)}; \theta)
\]

where \( P(y|x; \theta) \) is the neural translation model and \( \theta \) is the model parameter.

One big problem of the model training is that the history of any target word is correct and has been observed in the training data, but during inference, all the target words are predicted and may contain mistakes, which are fed as inputs to the model and quickly accumulated along with the sequence generation. This is called the exposure bias problem (Bengio et al. 2015).

### Our Approach

To deal with the exposure bias problem, we try to maximize the agreement between translations from L2R and R2L NMT models, and divide the NMT training objective into two parts: the standard maximum likelihood of training data, and the regularization terms that indicate the divergence of L2R and R2L models based on the current model parameter. In this section, we will start with basic model notations, followed by discussions of model regularization terms and efficient gradient approximation methods. In the last part, we show that the L2R and R2L NMT models can be jointly improved to achieve even better results.

### Notations

Given source sentence \( x = (x_1, x_2, \ldots, x_T) \) and its target translation \( y = (y_1, y_2, \ldots, y_T) \), let \( P(y|x; \theta) \) and \( P(y|x; \bar{\theta}) \) be L2R and R2L translation models, in which \( \theta \) and \( \bar{\theta} \) are corresponding model parameters. Specifically, L2R translation model can be decomposed as \( P(y|x; \theta) = \prod_{t=1}^T P(y_t|x_t; \theta) \), which means L2R model adopts previous targets \( y_1, \ldots, y_{t-1} \) as history to predict the current target \( y_t \) at each step \( t \), while the R2L translation model can similarly be decomposed as \( P(y|x; \bar{\theta}) = \prod_{t=T}^1 P(y_t|x_{T-t+1}; \bar{\theta}) \) and employs later targets \( y_{T-t+1}, \ldots, y_T \) as history to predict current target \( y_t \) at each step \( t \).

### NMT Model Regularization

Since L2R and R2L models are different chain decompositions of the same translation probability, output probabilities of the two models should be identical:

\[
\log P(y|x; \bar{\theta}) = \sum_{t=1}^{T} \log P(y_t|x_{<t}; \bar{\theta}) = \log P(y|x; \theta)
\]

However, if these two models are optimized separately by maximum likelihood estimation (MLE), there is no guarantee that the above equation will hold. To satisfy this constraint, we introduce two Kullbach-Leibler (KL) divergence regularization terms into the MLE training objective (Equation 1). For L2R model, the new training objective is:

\[
L(\theta) = \sum_{n=1}^N \log P(y^{(n)}|x^{(n)}; \theta) - \lambda \sum_{n=1}^N \text{KL}(P(y|x^{(n)}; \theta) || P(y|x^{(n)}; \bar{\theta})) - \lambda \sum_{n=1}^N \text{KL}(P(y|x^{(n)}; \bar{\theta}) || P(y|x^{(n)}; \theta))
\]

where \( \lambda \) is a hyper-parameter for regularization terms. These regularization terms are 0 when Equation 2 holds, otherwise
The definition of KL divergence, we have:

\[ KL(P(y|x^{(n)}; \theta) || P(y|x^{(n)}; \theta)) = \sum_{y \in Y(x^{(n)})} P(y|x^{(n)}; \theta) \log \frac{P(y|x^{(n)}; \theta)}{P(y|x^{(n)}; \theta)} \]

where \( Y(x^{(n)}) \) is a set of all possible candidate translations for the source sentence \( x^{(n)} \). Since \( P(y|x^{(n)}; \theta) \) is irrelevant to parameter \( \theta \), the partial derivative of this KL divergence with respect to \( \theta \) can be written as:

\[
\frac{\partial KL(P(y|x^{(n)}; \theta) || P(y|x^{(n)}; \theta))}{\partial \theta} = - \sum_{y \in Y(x^{(n)})} P(y|x^{(n)}; \theta) \frac{\partial \log P(y|x^{(n)}; \theta)}{\partial \theta}
\]

in which \( \frac{\partial \log P(y|x^{(n)}; \theta)}{\partial \theta} \) are the gradients specified with a standard sequence-to-sequence NMT network. The expectation \( E_{y \sim P(y|x^{(n)}; \theta)} \) can be approximated with samples from the R2L model \( P(y|x^{(n)}; \theta) \). Therefore, minimizing this regularization term is equivalent to maximizing the log-likelihood on the pseudo sentence pairs sampled from the R2L model.

For \( KL(P(y|x^{(n)}; \theta) || P(y|x^{(n)}; \theta)) \), similarly we have:

\[
KL(P(y|x^{(n)}; \theta) || P(y|x^{(n)}; \theta)) = \sum_{y \in Y(x^{(n)})} P(y|x^{(n)}; \theta) \log \frac{P(y|x^{(n)}; \theta)}{P(y|x^{(n)}; \theta)}
\]

The partial derivative of this KL divergence with respect to \( \theta \) is calculated as follows:

\[
\frac{\partial KL(P(y|x^{(n)}; \theta) || P(y|x^{(n)}; \theta))}{\partial \theta} = - E_{y \sim P(y|x^{(n)}; \theta)} \left( \log \frac{P(y|x^{(n)}; \theta)}{P(y|x^{(n)}; \theta)} \frac{\partial \log P(y|x^{(n)}; \theta)}{\partial \theta} \right)
\]

Similarly, we use sampling for the calculation of expectation \( E_{y \sim P(y|x^{(n)}; \theta)} \). There are two differences in Equation 7 compared with Equation 5: 1) Pseudo sentence pairs are not sampled from the R2L model \( P(y|x^{(n)}; \theta) \), but from the L2R model itself \( P(y|x^{(n)}; \theta) \); 2) \( \log \frac{P(y|x^{(n)}; \theta)}{P(y|x^{(n)}; \theta)} \) is used as weight to penalize incorrect pseudo pairs.

To sum up, the partial derivative of objective function \( L(\theta) \) with respect to \( \theta \) can be approximately written as follows:

\[
\frac{\partial L(\theta)}{\partial \theta} = \sum_{n=1}^{N} \frac{\partial \log P(y^{(n)}|x^{(n)}; \theta)}{\partial \theta} + \lambda \sum_{n=1}^{N} \sum_{y \sim P(y|x^{(n)}; \theta)} \frac{\partial \log P(y|x^{(n)}; \theta)}{\partial \theta} \]

The overall training is shown in Algorithm 1.

### Algorithm 1 Training Algorithm for L2R Model

**Input:** Bilingual Data \( D = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N} \)

**R2L Model** \( P(y|x; \theta) \)

**Output:** L2R Model \( P(y|x; \theta) \)

1: **procedure** TRAINING PROCESS
2: **while** Not Converged **do**
3: Sample sentence pairs \( (x^{(n)}, y^{(n)}) \) from bilingual data \( D \);
4: Generate \( m \) translation candidates \( y_1^{(n)}, \ldots, y_m^{(n)} \) for \( x^{(n)} \) by translation model \( P(y|x^{(n)}; \theta) \) and build pseudo translation candidates \( \{y_1^{(n)}, y_1^{(n)}\}_{i=1}^{m} \);
5: Generate \( m \) translation candidates \( y_1^{(n)}, \ldots, y_m^{(n)} \) for \( x^{(n)} \) by translation model \( P(y|x^{(n)}; \theta) \) and build pseudo translation candidates \( \{y_1^{(n)}, y_1^{(n)}\}_{i=1}^{m} \) weighted with \( P(y_1^{(n)}|x^{(n)}; \theta) \);
6: Update \( P(y|x; \theta) \) with Equation 8 given original data \( \{(x^{(n)}, y^{(n)})\}_{n=1}^{N} \) and two synthetic data \( \{(x^{(n)}, y_1^{(n)})\}_{i=1}^{m} \) and \( \{(x^{(n)}, y_1^{(n)})\}_{i=1}^{m} \).
7: **end while**
8: **end procedure**

**Joint Training for Paired NMT Models**

In practice, due to the imperfection of R2L model, the agreement between L2R and R2L models sometimes may mislead L2R model training. On the other hand, due to the symmetry of L2R and R2L models, L2R model can also serve as the discriminator to punish bad translation candidates generated from R2L model. Similarly, the objective function of
models with an iterative process. In each iteration, we fix R2L model and use it as a helper to optimize L2R model with Equation 3, and at the same time, we fix L2R model and use it as a helper to optimize R2L model with Equation 9. The iterative training continues until the performance on development set does not increase.

**Experiments**

**Setup**

To examine the effectiveness of our proposed approach, we conduct experiments on three datasets, including NIST OpenMT for Chinese-English, WMT17 for English-German and Chinese-English. In all experiments, we use BLEU (Papineni et al. 2002) as the automatic metric for translation evaluation.

**Datasets.** For NIST OpenMT’s Chinese-English translation task, we select our training data from LDC corpora,\(^1\) which consists of 2.6M sentence pairs with 65.1M Chinese words and 67.1M English words respectively. Any sentence longer than 80 words is removed from training data. The NIST OpenMT 2006 evaluation set is used as validation set, and NIST 2003, 2005, 2008, 2012 datasets as test sets. We limit the vocabulary to contain up to 50K most frequent words on both source and target sides, and convert remaining words into the \(<\text{unk}>\) tokens. In decoding period, we follow Luong et al. (2015) to handle the \(<\text{unk}>\) replacement.

For WMT17’s English-German translation task, we use the pre-processed training data provided by the task organizers.\(^2\) The training data consists of 5.8M sentence pairs with 141M English words and 134M German words respectively. We use the newsdev2016 as the validation set and the newstest2017 as the test set. The maximal sentence length is set as 128. For vocabulary, we use 37K sub-word tokens based on Byte Pair Encoding (BPE) (Sennrich, Haddow, and Birch 2016b).

For WMT17’s Chinese-English translation task, we use all the available parallel data, which consists of 24M sentence pairs, including News Commentary, UN Parallel Corpus and CWMT Corpus.\(^3\) The newsdev2017 is used as the validation set and newstest2017 as the test set. We also limit the maximal sentence length to 128. For data pre-processing, we segment Chinese sentences with our in-house Chinese word segmentation tool and tokenize English sentences with the scripts provided in Moses.\(^4\) Then we learn a BPE model on pre-processed sentences with 32K merge operations, in which 44K and 33K sub-word tokens are adopted as source and target vocabularies respectively.

---

\(^1\)The corpora include LDC2002E17, LDC2002E18, LDC2003E07, LDC2003E14, LDC2005E83, LDC2005T06, LDC2006T10, LDC2006E17, LDC2006E26, LDC2006E34, LDC2006E85, LDC2006E92, LDC2006T06, LDC2004T08, LDC2005T10

\(^2\)http://data.statmt.org/wmt17/translation-task/preprocessed/de-en/

\(^3\)http://www.statmt.org/wmt17/translation-task.html

\(^4\)https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl
Table 2 shows the evaluation results of different models on NIST datasets. MRT represents Shen et al. (2016)’s method, RT denotes our regularization approach, and JS represents Liu et al. (2016)’s method that modifies the inference strategy by reranking the N-best results with a joint probability of KL divergence. In practice, to speed up the decoding process, we sort all source sentences according to the sentence length, and then 32 sentences are simultaneously translated with parallel decoding implementation. In our experiments, we try different settings (λ = 0.1, 0.2, 0.5, 2.0, 5.0), and find the λ = 1 achieves the best BLEU result on validation set. We also test the model performance on validation set with the bigger parameter m. When m = 2, 3, 4, we do not find the further improvement but it brings some training times due to more pseudo sentence pairs. In addition, we use sentence-level BLEU to filter wrong translations whose BLEU score is not greater than 30%. Notice that the R2L model gets comparable results with the L2R model, thus only the result of the L2R model is reported in our experiments.

Evaluation on NIST Corpora

Table 2 shows the evaluation results of different models on NIST datasets. MRT represents Shen et al. (2016)’s method, RT denotes our regularization approach, and JS represents Liu et al. (2016)’s method that modifies the inference strategy by reranking the N-best results with a joint probability of bidirectional models. All the results are reported based on case-insensitive BLEU and computed using Moses multi-bleu.perl script.

We observe that by taking agreement information into consideration, Transformer+JS and Transformer+RT can bring improvement across different test sets, in which our approach achieves 2.73 BLEU point improvements over Transformer on average. These results confirm that introducing agreement between L2R and R2L models helps handle exposure bias problem and improves translation quality.

Besides, we see that Transformer+RT gains better performance than Transformer+JS across different test sets, with 1.5 BLEU point improvements on average. Since Transformer+JS only leverages the agreement restriction in inference stage, L2R and R2L models still suffer from exposure bias problem and generate bad translation candidates, which limits the room for improvement in the re-ranking process. Instead of combining R2L model during inference, our approach utilizes the intrinsic probabilistic connection between L2R and R2L models to guide the learning process. The two NMT models are expected to adjust in disagreement cases and then the exposure bias problem of them can be solved.

Longer source sentence implies longer translation that more easily suffers from exposure bias problem. To further verify our approach, we group source sentences of similar length together and calculate the BLEU score for each group. As shown in Figure 2, we can see that our mechanism achieves the best performance in all groups. The gap between our method and the other three methods is small when the length is smaller than 10, and the gap becomes bigger when the sentences become longer. This further confirms the efficiency of our proposed method in dealing with the explore bias problem.
**Table 3:** Case-sensitive BLEU scores (%) for English-German and Chinese-English translation on WMT test sets. Edinburgh (Sennrich et al. 2017) and Sogou (Wang et al. 2017b) NMT systems are No.1 system in leaderboard of WMT 2017’s English-German and Chinese-English translation tasks respectively.

| System                        | English-German | Chinese-English |
|-------------------------------|----------------|-----------------|
|                              | newstest2016   | newstest2017    | newsdev2017    | newstest2017 |
| Transformer                   | 32.58          | 25.48           | 20.87          | 23.01        |
| Transformer+MRT               | 33.27          | 25.87           | 21.66          | 24.24        |
| Transformer+JS                | 32.91          | 25.93           | 21.25          | 23.59        |
| Transformer+RT                | 34.56          | 27.18           | 22.50          | 25.38        |
| Transformer-big               | 33.58          | 27.13           | 21.91          | 24.03        |
| Transformer-big+BT            | 35.06          | 28.34           | 23.59          | 25.53        |
| Transformer-big+BT+RT         | **36.78**      | **29.46**       | **24.84**      | **27.21**    |
| Edinburgh’s NMT System (ensemble) | 36.20          | 28.30           | 24.00          | 25.70        |
| Sogou’s NMT System (ensemble) | -              | -               | 22.90          | 26.40        |

Table 4: Translation performance of our method on WMT validation sets during training process. “Iteration 0” denotes baseline Transformer model.

|                      | English-German | Chinese-English |
|----------------------|----------------|-----------------|
| Iteration 0          | 32.58          | 20.87           |
| Iteration 1          | 33.86          | 21.92           |
| Iteration 2          | 34.56          | 22.50           |
| Iteration 3          | 34.58          | 22.47           |

**Evaluation on WMT17 Corpora**

For WMT17 Corpora, we verify the effectiveness of our approach on English-German and Chinese-English translation tasks from two angles: 1) We compare our approach with baseline systems when only parallel corpora is used; 2) We investigate the impact of the combination of back-translation technique (Sennrich, Haddow, and Birch 2016a) and our approach. Since back-translation method brings in more synthetic data, we choose Transformer-big settings defined in Vaswani et al. (2017) for this setting. Experimental results are shown in Table 3, in which BT denotes back-translation method. In order to be comparable with NMT systems reported in WMT17, all results are reported based on case-sensitive BLEU and computed using official tools - SacreBLEU.  

**Only Parallel Corpora.** As shown in Table 3, Transformer+JS and Transformer+RT both generate improvements on test sets for English-German and Chinese-English translation tasks. It confirms the effectiveness of leveraging agreements between L2R and R2L models. Additionally, our approach significantly outperforms Transformer+JS, yielding the best BLEU score on English-German and Chinese-English test sets respectively when merely using bilingual data. These results further prove the effectiveness of our method.

**Combining with Back-Translation Method.** To verify the effect of our approach when monolingual data is available, we combine our approach with back-translation method. We first randomly select 5M German sentences and 12M English sentences from “News Crawl: articles from 2016”. Then German-English and English-Chinese NMT systems learnt with parallel corpora are used to translate monolingual target sentences.

From Table 3, we find that Transformer-big gains better performance than Transformer due to more model parameters. When back-translation method is employed, Transformer-big+BT achieves 1.21 and 1.5 BLEU point improvements over Transformer-big on English-German and Chinese-English sets respectively. Our method can further gain remarkable improvements based on back-translation method, resulting in the best results on all translation tasks. These results show that in semi-supervised learning scenarios, the NMT model still can benefit from our proposed approach. In addition, our single model Transformer-big+BT+RT even achieves the best performance on WMT17’s English-German and Chinese-English translation tasks, among all the reported results, including ensemble systems.

**Effect of Joint Training**

We further investigate the impact of our joint training algorithm during the whole training process. Table 4 shows the BLEU scores on WMT validation sets in each iteration. For each iteration, we train NMT models until the performance on the development set does not increase. We find that more iterations can lead to better evaluation results consistently and 2 iterations are enough to reach convergence in our experiments. However, further iterations cannot bring noticeable translation accuracy improvements but more training time. For training cost, since our method is based on pre-trained models, the entire training time is almost 2 times the original MLE training.

**Example**

In this section, we give a case study to analyze our method. Table 5 provides a Chinese-English translation example.
The victim's brother, Louis Galia, told ABC's station KGO, "age desirable suffixes from Transformer (R2L). punish bad suffixes generated by Transformer and encourage between L2R and R2L models in training stage can better improve the agreement of the former (R2L). The reason is that leveraging the agreement between L2R and R2L NMT models in training stage can better punish bad suffixes generated by Transformer and encourage desirable suffixes from Transformer (R2L).

### Related Work

Target-bidirectional transduction techniques have been explored in statistical machine translation, under the IBM framework (Watanabe and Sumita 2002) and the feature-driven linear models (Finch and Sumita 2009; Zhang et al. 2013). Recently, Liu et al. (2016) and Zhang et al. (2018) migrate this method from SMT to NMT by modifying the inference strategy and decoder architecture of NMT. Liu et al. (2016) propose to generate N-best translation candidates from L2R and R2L NMT models and leverage the joint probability of two models to find the best candidates from the combined N-best list. Zhang et al. (2018) design a two-stage decoder architecture for NMT, which generates translation candidates in a right-to-left manner in first-stage and then gets final translation based on source sentence and previous generated R2L translation. Different from their method, our approach directly exploits the target-bidirectional agreement in training stage by introducing regularization terms. Without changing the neural network architecture and inference strategy, our method keeps the same speed as the original model during inference.

To handle the exposure bias problem, many methods have been proposed, including designing new training objectives (Shen et al. 2016; Wiseman and Rush 2016) and adopting reinforcement learning approaches (Ranzato et al. 2015; Bahdanau et al. 2016). Shen et al. (2016) attempt to directly minimize expected loss (maximize the expected BLEU) with Minimum Risk Training (MRT). Wiseman and Rush (2016) adopt a beam-search optimization algorithm to reduce inconsistency between training and inference. Besides, Ranzato et al. (2015) propose a mixture training method to perform a gradual transition from MLE training to BLEU score optimization using reinforcement learning. Bahdanau et al. (2016) design an actor-critic algorithm for sequence prediction, in which the NMT system is the actor, and a critic network is proposed to predict the value of output tokens. Instead of designing task-specific objective functions or complex training strategies, our approach only adds regularization terms to standard training objective function, which is simple to implement but effective.

### Conclusion

In this paper, we have presented a simple and efficient regularization approach to neural machine translation, which relies on the agreement between L2R and R2L NMT models. In our method, two Kullback-Leibler divergences based on probability distributions of L2R and R2L models are added to the standard training objective as regularization terms. An efficient approximation algorithm is designed to enable fast training of the regularized training objective and then a training strategy is proposed to jointly optimize L2R and R2L models. Empirical evaluations are conducted on Chinese-English and English-German translation tasks, demonstrating that our approach leads to significant improvements compared with strong baseline systems.

In our future work, we plan to test our method on other sequence-to-sequence tasks, such as summarization and dialogue generation. Besides the back-translation method, it is also worth trying to integrate our approach with other semi-supervised methods to better leverage unlabeled data.

### Acknowledgments

This research was partially supported by grants from the National Key Research and Development Program of China (Grant No. 2018YFB1004300), the National Natural Science Foundation of China (Grant No. 61703386), the Anhui Provincial Natural Science Foundation (Grant No. 1708085QF140), and the Fundamental Research Funds for the Central Universities (Grant No. WK2150110006).

Besides, we appreciate Dongdong Zhang and Ren Shuo for the fruitful discussions. We also thank the anonymous
reviewers for their careful reading of our paper and insightful comments.

References
Bahdanau, D.; Brakel, P.; Xu, K.; Goyal, A.; Lowe, R.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2016. An actor-critic algorithm for sequence prediction. CoRR abs/1607.07086.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473.

Bengio, S.; Vinyals, O.; Jaitly, N.; and Shazeer, N. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In NIPS.

Chiang, D. 2007. Hierarchical phrase-based translation. Computational Linguistics.

Cho, K.; van Merrienboer, B.; Çaglar Gulçehre; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In EMNLP.

Eriguchi, A.; Hashimoto, K.; and Tsuruoka, Y. 2016. Tree-to-sequence attentional neural machine translation. In ACL.

Finch, A. M., and Sumita, E. 2009. Bidirectional phrase-based neural machine translation. In ACL.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. 2017. Convolutional sequence to sequence learning. In ICML.

Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In AISTATS.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. Neural computation.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. CoRR abs/1412.6980.

Koehn, P.; Och, F. J.; and Marcu, D. 2003. Statistical phrase-based translation. In HLT-NAACL.

Liu, L.; Utiyama, M.; Finch, A. M.; and Sumita, E. 2016. Agreement on target-bidirectional neural machine translation. In HLT-NAACL.

Luong, T.; Sutskever, I.; Le, Q. V.; Vinyals, O.; and Zaremba, W. 2015. Addressing the rare word problem in neural machine translation. In ACL.

Papineni, K.; Roucos, S. E.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In ACL.

Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2015. Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732.

Sennrich, R., Birch, A.; Currey, A.; Germann, U.; Haddow, B.; Heathfield, K.; Barone, A. V.M.; and Williams, P. G. 2017. The university of edinburgh’s neural mt systems for wmt17. In WMT.

Sennrich, R.; Haddow, B.; and Birch, A. 2016a. Improving neural machine translation models with monolingual data. In ACL.

Sennrich, R.; Haddow, B.; and Birch, A. 2016b. Neural machine translation of rare words with subword units. In ACL.

Shen, S.; Cheng, Y.; He, Z.; He, W.; Wu, H.; Sun, M.; and Liu, Y. 2016. Minimum risk training for neural machine translation. In ACL.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In NIPS.

Tu, Z.; Lu, Z.; Liu, Y.; Liu, X.; and Li, H. 2016. Modeling coverage for neural machine translation. In ACL.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In NIPS.

Wang, M.; Lu, Z.; Zhou, J.; and Liu, Q. 2017a. Deep neural machine translation with linear associative unit. In ACL.

Wang, Y.; Cheng, S.; Jiang, L.; Yang, J.; Chen, W.; Li, M.; Shi, L.; Wang, Y.; and Yang, H. 2017b. Sogou neural machine translation systems for wmt17. In WMT.

Watanabe, T., and Sumita, E. 2002. Bidirectional decoding for statistical machine translation. In COLING.

Wiseman, S., and Rush, A. M. 2016. Sequence-to-sequence learning as beam-search optimization. In EMNLP.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kuryan, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G. S.; Hughes, M.; and Dean, J. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. CoRR abs/1609.08144.

Zhang, H.; Toutanova, K.; Quirk, C.; and Gao, J. 2013. Beyond left-to-right: Multiple decomposition structures for smt. In HLT-NAACL.

Zhang, X.; Su, J.; Qin, Y.; Liu, Y.; Ji, R.; and Wang, H. 2018. Asynchronous bidirectional decoding for neural machine translation. In AAAI.