

English-Japanese Neural Machine Translation with Encoder-Decoder-Reconstructor

Yukio Matsumura

Takayuki Sato

Mamoru Komachi

Tokyo Metropolitan University

Tokyo, Japan

matsumura-yukio@ed.tmu.ac.jp, sasatatata99@gmail.com, komachi@tmu.ac.jp

Abstract

Neural machine translation (NMT) has recently become popular in the field of machine translation. However, NMT suffers from the problem of repeating or missing words in the translation. To address this problem, [Tu et al. \(2017\)](#) proposed an encoder-decoder-reconstructor framework for NMT using back-translation. In this method, they selected the best forward translation model in the same manner as [Bahdanau et al. \(2015\)](#), and then trained a bi-directional translation model as fine-tuning. Their experiments show that it offers significant improvement in BLEU scores in Chinese-English translation task. We confirm that our re-implementation also shows the same tendency and alleviates the problem of repeating and missing words in the translation on a English-Japanese task too. In addition, we evaluate the effectiveness of pre-training by comparing it with a jointly-trained model of forward translation and back-translation.

1 Introduction

Recently, neural machine translation (NMT) has gained popularity in the field of machine translation. The conventional encoder-decoder NMT proposed by [Cho et al. \(2014\)](#) uses two recurrent neural networks (RNN): one is an encoder, which encodes a source sequence into a fixed-length vector, and the other is a decoder, which decodes the vector into a target sequence. A newly proposed attention-based NMT by [Bahdanau et al. \(2015\)](#) can predict output words using the weights of each hidden state of the encoder by the attention mechanism, improving the adequacy of translation.

Even with the success of attention-based models, a number of open questions remain in NMT. [Tu et al. \(2016\)](#) argued two of the common problems are over-translation: some words are repeatedly translated unnecessary and under-translation: some words are mistakenly untranslated. This is due to the fact that NMT can not completely convert the information from the source sentence to the target sentence. [Mi et al. \(2016\)](#) and [Feng et al. \(2016\)](#) pointed out that NMT lacks the notion of coverage vector in phrase-based statistical machine translation (PBSMT), so unless otherwise specified, there is no way to prevent missing translations.

Another problem in NMT is an objective function. NMT is optimized by cross-entropy; therefore, it does not directly maximize the translation accuracy. [Shen et al. \(2016\)](#) pointed out that optimization by cross-entropy is not appropriate and proposed a method of optimization based on a translation accuracy score, such as expected BLEU, which led to improvement of translation accuracy. However, BLEU is an evaluation metric based on n-gram precision; therefore, repetition of some words may be present in the translation even though the BLEU score is improved.

To address problem of repeating and missing words in the translation, [Tu et al. \(2017\)](#) introduce an encoder-decoder-reconstructor framework that optimizes NMT by back-translation from the output sentences into the original source sentences. In their method, after training the forward translation in a manner similar to the conventional attention-based NMT, they train a back-translation model from the hidden state of the decoder into the source sequence by a new decoder to enforce agreement between source and target sentences.

In order to confirm the language independence of the framework, we experiment on two parallel corpora of English-Japanese and Japanese-

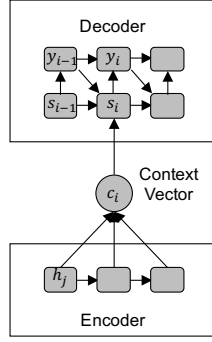


Figure 1: Attention-based NMT.

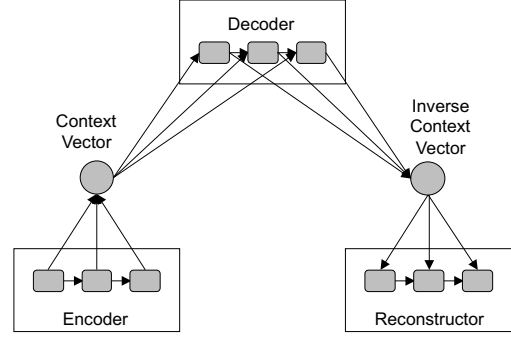


Figure 2: Encoder-Decoder-Reconstructor.

English translation tasks using encode-decoder-reconstructor. Our experiments show that their method offers significant improvement in BLEU scores and alleviates the problem of repeating and missing words in the translation on English-Japanese translation task, though the difference is not significant on Japanese-English translation task.

In addition, we jointly train a model of forward translation and back-translation without pre-training, and then evaluate this model. As a result, the encoder-decoder-reconstructor can not be trained well without pre-training, so it proves that we have to train the forward translation model in a manner similar to the conventional attention-based NMT as pre-training.

The main contributions of this paper are as follows:

- Experimental results show that encode-decoder-reconstructor framework achieves significant improvements in BLEU scores (1.0-1.4) for English-Japanese translation task.
- Experimental results show that encode-decoder-reconstructor framework has to train the forward translation model in a manner similar to the conventional attention-based NMT as pre-training.

2 Related Works

Several studies have addressed the NMT-specific problem of missing or repeating words. Niehues et al. (2016) optimized NMT by adding the outputs of PBSMT to the input of NMT. Mi et al. (2016) and Feng et al. (2016) introduced a distributed version of coverage vector taken from PBSMT to consider which words have been already

translated. All these methods, including ours, employ information of the source sentence to improve the quality of translation, but our method uses back-translation to ensure that there is no inconsistency. Unlike other methods, once learned, our method is identical to the conventional NMT model, so it does not need any additional parameters such as coverage vector or a PBSMT system for testing.

The attention mechanism proposed by Meng et al. (2016) considers not only the hidden states of the encoder but also the hidden states of the decoder so that over-translation can be relaxed. In addition, the attention mechanism proposed by Feng et al. (2016) computes a context vector by considering the previous context vector to prevent over-translation. These works indirectly reduce repeating and missing words, while we directly penalize translation mismatch by considering back-translation.

The encoder-decoder-reconstructor framework for NMT proposed by Tu et al. (2017) optimizes NMT by reconstructor using back-translation. They consider likelihood of both of forward translation and back-translation, and then this framework offers significant improvement in BLEU scores and alleviates the problem of repeating and missing words in the translation on a Chinese-English translation task.

3 Neural Machine Translation

Here, we describe the attention-based NMT proposed by Bahdanau et al. (2015) as shown in Figure 1.

The input sequence ($\mathbf{x} = [x_1, x_2, \dots, x_{|\mathbf{x}|}]$) is converted into a fixed-length vector by the encoder using an RNN. At each time step t , the hidden state

h_t of the encoder is presented as

$$h_t = [\vec{h}_t^\top : \overleftarrow{h}_t^\top]^\top \quad (1)$$

using a bidirectional RNN. The forward state \vec{h}_t and the backward state \overleftarrow{h}_t are computed by

$$\vec{h}_t = r(x_t, h_{t-1}) \quad (2)$$

and

$$\overleftarrow{h}_t = r'(x_t, h_{t+1}) \quad (3)$$

where r and r' are nonlinear functions. The hidden states $(h_1, h_2, \dots, h_{|x|})$ are converted into a fixed-length vector v as

$$v = q([h_1, h_2, \dots, h_{|x|}]) \quad (4)$$

where q is a nonlinear function.

The fixed-length vector v generated by the encoder is converted into the target sequence ($\mathbf{y} = [y_1, y_2, \dots, y_{|y|}]$) by the decoder using an RNN. At each time step i , the conditional probability of the output word \hat{y}_i is computed by

$$p(\hat{y}_i | \mathbf{y}_{<i}, \mathbf{x}) = f(s_i, y_{i-1}, c_i) \quad (5)$$

where f is a nonlinear function. The hidden state s_i of the decoder is presented as

$$s_i = g(s_{i-1}, y_{i-1}, c_i) \quad (6)$$

using the hidden state s_{i-1} and the target word y_{i-1} at the previous time step and the context vector c_i .

The context vector c_i is a weighted sum of each hidden state h_j of the encoder. It is presented as

$$c_i = \sum_{j=1}^{|x|} \alpha_{ij} h_j \quad (7)$$

and its weight α_{ij} is a normalized probability distribution. It is computed by

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{|x|} \exp(e_{ik})} \quad (8)$$

and

$$e_{ij} = v_a^\top \tanh(W_a s_{i-1} + U_a h_j) \quad (9)$$

where v_a is a weight vector and W_a and U_a are weight matrices.

	ASPEC	NTCIR
train	827,188	1,169,201
dev	1,504	2,741
test	1,556	2,300

Table 1: Numbers of parallel sentences.

The objective function is defined by

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^{|y|} \log p(\hat{y}_i^{(n)} | \mathbf{y}_{<i}^{(n)}, \mathbf{x}^{(n)}, \theta) \quad (10)$$

where N is the number of data and θ is a model parameter.

Incidentally, as a nonlinear function, the hyperbolic tangent function or the rectified linear unit are generally used.

4 Encoder-Decoder-Reconstructor

4.1 Architecture

Next, we describe the encoder-decoder-reconstructor framework for NMT proposed by Tu et al. (2017) as shown in Figure 2. The encoder-decoder-reconstructor consists of two components: the standard encoder-decoder as an attention-based NMT proposed by Bahdanau et al. (2015) and the **reconstructor** which back-translates from the hidden states of decoder to the source sentence.

In their method, the hidden state of the decoder is back-translated into the source sequence (\mathbf{x}) by the reconstructor for the back-translation. At each time step i , the conditional probability of the output word \hat{x}_i is computed by

$$p(\hat{x}_i | \mathbf{x}_{<i}, \hat{\mathbf{y}}) = f'(s'_i, x_{i-1}, c'_i) \quad (11)$$

where f' is a nonlinear function. The hidden state s'_i of the reconstructor is presented as

$$s'_i = g'(s'_{i-1}, x_{i-1}, c'_i) \quad (12)$$

using the hidden state s'_{i-1} and the source word x_{i-1} at the previous time step and the new context vector (**inverse context vector**) c'_i .

The inverse context vector c'_i is a weighted sum of each hidden state s_j of the decoder (on forward translation). It is presented as

$$c'_i = \sum_{j=1}^{|y|} \alpha'_{ij} s_j \quad (13)$$

English-Japanese				
Corpus	Model	BLEU	<i>p</i> -value	Hours
ASPEC	Baseline-NMT	29.75	-	99
	+Reconstructor	30.76	0.00	149
	+Reconstructor (Jointly-Training)	26.04	-	174
NTCIR	Baseline-NMT	30.03	-	116
	+Reconstructor	31.40	0.00	166
	+Reconstructor (Jointly-Training)	29.04	-	252

Table 2: English-Japanese translation results.

Japanese-English				
Corpus	Model	BLEU	<i>p</i> -value	Hours
ASPEC	Baseline-NMT	21.91	-	87
	+Reconstructor	22.27	0.10	127
	+Reconstructor (Jointly-Training)	16.29	-	187
NTCIR	Baseline-NMT	29.48	-	180
	+Reconstructor	29.73	0.11	244
	+Reconstructor (Jointly-Training)	28.95	-	300

Table 3: Japanese-English translation results.

and its weight α'_{ij} is a normalized probability distribution. It is computed by

$$\alpha'_{ij} = \frac{\exp(e'_{ij})}{\sum_{k=1}^{|y|} \exp(e'_{ik})} \quad (14)$$

and

$$e'_{ij} = v'_a \top \tanh(W'_a s'_{i-1} + U'_a s_j) \quad (15)$$

where v'_a is a weight vector and W'_a and U'_a are weight matrices.

The objective function is defined by

$$\begin{aligned} \mathcal{L}(\theta, \gamma) = & \frac{1}{N} \sum_{n=1}^N \left\{ \sum_{i=1}^{|y|} \log p(\hat{y}_i^{(n)} | \mathbf{y}_{<i}^{(n)}, \mathbf{x}^{(n)}, \theta) \right. \\ & \left. + \lambda \sum_{i=1}^{|x|} \log p(\hat{x}_i^{(n)} | \mathbf{x}_{<i}^{(n)}, \mathbf{s}^{(n)}, \gamma) \right\} \end{aligned} \quad (16)$$

where N is the number of data, θ and γ are model parameters and λ is a hyper-parameter which can consider the weight between forward translation and back-translation.

This objective function consists of two parts: forward measures translation fluency, and backward measures translation adequacy. Thus, the combined objective function is more consistent with the goal of enhancing overall translation quality, and can more effectively guide the parameter training for making better translation.

4.2 Training

The encoder-decoder-reconstructor is trained with likelihood of both the encoder-decoder and the reconstructor on a set of training datasets. [Tu et al. \(2017\)](#) trained a back-translation model from the hidden state of the decoder into the source sequence by reconstructor to enforce agreement between source and target sentences using Equation 16 after training the forward translation in a manner similar to the conventional attention-based NMT using Equation 10.

In addition, we experiment to jointly train a model of forward translation and back-translation without pre-training. It may learn a globally optimal model compared to locally optimal model pre-trained using the forward translation.

4.3 Testing

[Tu et al. \(2017\)](#) used a beam search to predict target sentences that approximately maximizes both of forward translation and back-translation on testing. In this paper, however, we do not use a beam search for simplicity and effectiveness.

5 Experiments

We evaluated the encoder-decoder-reconstructor framework for NMT on English-Japanese and Japanese-English translation tasks.

Example 1: Improvement in under-translation.	
Input	the conditions under which the effect of turbulent viscosity is correctly evaluated were examined <u>on the basis of the relation between turbulent viscosity and numerical viscosity in size</u> .
Baseline-NMT	乱流粘性の影響を正確に評価する条件を検討した。
+Reconstructor	乱流粘性の影響を正確に評価する条件を, <u>乱流粘性と数値的粘性の関係を基に調べた。</u>
+Reconstructor (Jointly-Training)	乱流粘性の影響を考慮した条件を, <u>乱流粘性と粘性の粘性との関係をもとに検討した。</u>
Reference	<u>乱流粘性と数値粘性の大小関係により</u> ,乱流粘性の効果が正しく評価される条件を検討した。
Example 2: Improvement in over-translation.	
Input	activity was high in cells of the young , especially <u>newborn infant</u> , and was very slight in cells of <u>30 - year - old or more</u> .
Baseline-NMT	活動性は若齢,特に <u>新生児</u> では <u>30歳以上の細胞</u> で高く, <u>30歳以上の細胞</u> ではわずかであった。
+Reconstructor	その活性は若齢,特に <u>新生児</u> は細胞が高く, <u>30歳以上の細胞</u> ではわずかであった。
+Reconstructor (Jointly-Training)	若齢の <u>新生児</u> では活性は高かったが, <u>30歳以上</u> の場合には極めて軽度であった。
Reference	活性は若い個体,特に <u>新生児</u> の細胞で高く, <u>30歳以上</u> のものではごくわずかであった。

Table 4: Examples of outputs of English-Japanese translation.

5.1 Datasets

We used two parallel corpora: Asian Scientific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2016) and NTCIR PatentMT Parallel Corpus (Goto et al., 2013). Regarding the training data of ASPEC, we used only the first 1 million sentences sorted by sentence-alignment similarity. Japanese sentences were segmented by the morphological analyzer MeCab (version 0.996, IPADIC), and English sentences were tokenized by tokenizer.perl of Moses. Table 1 shows the numbers of the sentences in each corpus. Note that sentences with more than 40 words were excluded from the training data.

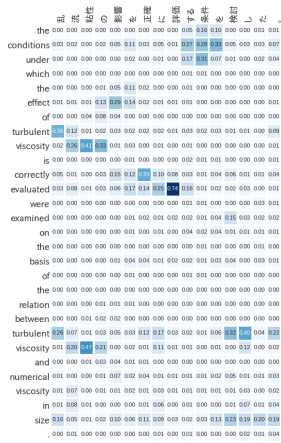
5.2 Models

We used the attention-based NMT (Bahdanau et al., 2015) as a baseline-NMT, the encoder-decoder-reconstructor (Tu et al., 2017) and the encoder-decoder-reconstructor that jointly trained forward translation and back-translation without pre-training. The RNN used in the experiments had 512 hidden units, 512 embedding units, 30,000 vocabulary size and 64 batch size. We

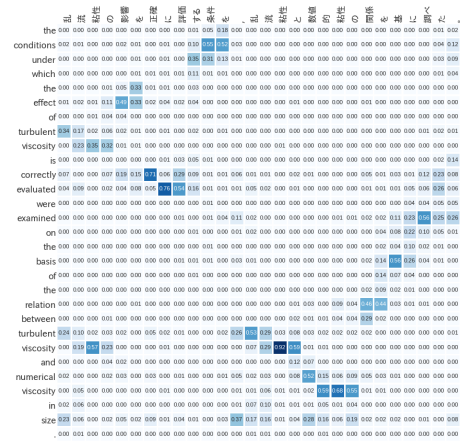
used Adagrad (initial learning rate 0.01) for optimizing model parameters. We trained our model on GeForce GTX TITAN X GPU. Note that we set the hyper-parameter $\lambda = 1$ on the encoder-decoder-reconstructor same as Tu et al. (2017).

5.3 Results

Tables 2 and 3 show the translation accuracy in BLEU scores, the p -value of the significance test by bootstrap resampling (Koehn, 2004) and training time in hours until convergence. The encoder-decoder-reconstructor (Tu et al., 2017) requires slightly longer time to train than the baseline NMT, but we emphasize that decoding time remains the same with the encoder-decoder-reconstructor and baseline-NMT. The results show that the encoder-decoder-reconstructor (Tu et al., 2017) significantly improves translation accuracy by 1.01 points on ASPEC and 1.37 points on NTCIR in English-Japanese translation ($p < 0.05$). However, it does not significantly improve translation accuracy in Japanese-English translation. In addition, it is proved that the encoder-decoder-reconstructor without pre-training worsens rather

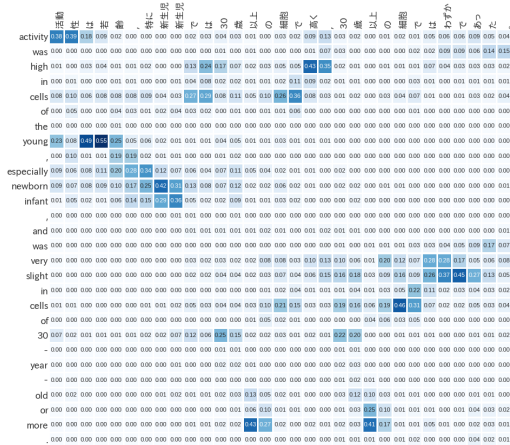


Baseline-NMT

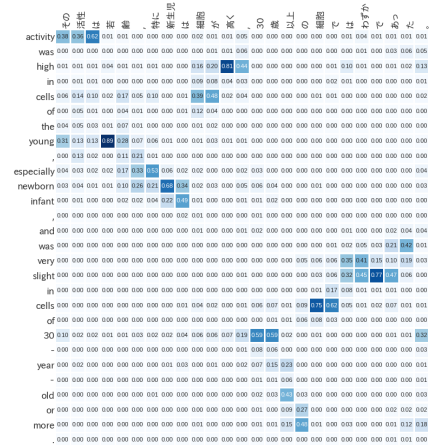


Encoder-Decoder-Reconstructor

Figure 3: The attention layer in Example 1 : Improvement in under-translation.



Baseline-NMT



Encoder-Decoder-Reconstructor

Figure 4: The attention layer in Example 2 : Improvement in over-translation.

than improves translation accuracy.

Table 4 shows examples of outputs of English-Japanese translations. In Example 1, “乱流粘性 と数値粘性の大小関係により,” (on the basis of the relation between turbulent viscosity and numerical viscosity in size) is missing in the output of baseline-NMT, but “乱流粘性 と数値的粘性の関係を基に” (on the basis of the relation between turbulent viscosity and numerical viscosity) is present in the output of encoder-decoder-reconstructor. In Example 2, “新生児” (newborn infant) and “30歳以上の” (of 30 - year - old or more) are repeated in the output of baseline-NMT, but they appear only once in the output of encoder-decoder-reconstructor.

In addition, Figures 3 and 4 show the attention layer on baseline-NMT and encoder-decoder-reconstructor in each example. In Figure 3, although the attention layer of baseline NMT attends input word “turbulent”, the decoder does not output “乱流” (turbulent) but “検討” (examined) at the 13th word. Thus, under-translation may be resulted from the hidden layer or the embedding layer instead of the attention layer. In Figure 4, it is found that the attention layer of baseline-NMT repeatedly attends input words “newborn infant” and “30 - year - old or more”. Consequently, the decoder repeatedly outputs “新生児” (newborn infant) and “30歳以上の” (of 30 - year - old or more). On the other hand, the attention layer

Corpus	Model	English-Japanese			Japanese-English		
		(i)	(ii)	(iii)	(i)	(ii)	(iii)
ASPEC	Baseline-NMT	1,141	378	1,045	951	494	1,085
	+Reconstructor	988	336	1,042	836	418	1,014
	+Reconstructor (Jointly-Training)	1,292	446	1,147	1,106	525	1,821
NTCIR	Baseline-NMT	2,122	1,015	1,106	2,521	1,073	1,630
	+Reconstructor	1,958	922	963	2,187	987	1,422
	+Reconstructor (Jointly-Training)	1,978	916	1,078	2,475	1,107	1,610

Table 5: Numbers of redundant and unknown word tokens.

of encoder-decoder-reconstructor almost correctly attends input words.

Table 5 shows a comparison of the number of word occurrences for each corpus and model. The columns show (i) the number of words that appear more frequently than the counterparts in the reference, and (ii) the number of words that appear more than once but are not included in the reference. Note that these numbers do not include unknown words, so (iii) shows the number of unknown words. In all the cases, the number of occurrence of redundant words is reduced in encoder-decoder-reconstructor. Thus, we confirmed that encoder-decoder-reconstructor achieves reduction of repeating and missing words while maintaining the quality of translation.

6 Conclusion

In this paper, we evaluated the encoder-decoder-reconstructor on English-Japanese and Japanese-English translation tasks. In addition, we evaluate the effectiveness of pre-training by comparing it with a jointly-trained model of forward translation and back-translation. Experimental results show that the encoder-decoder-reconstructor offers significant improvement in BLEU scores and alleviates the problem of repeating and missing words in the translation on English-Japanese translation task, and the encoder-decoder-reconstructor can not be trained well without pre-training, so it proves that we have to train the forward translation model in a manner similar to the conventional attention-based NMT as pre-training.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–15.
- 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. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.
- Shi Feng, Shujie Liu, Nan Yang, Mu Li, and Ming Zhou. 2016. Improving Attention Modeling with Implicit Distortion and Fertility for Machine Translation. *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 3082–3092.
- Isao Goto, Ka-Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K Tsou. 2013. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. *Proceedings of the 10th NII Testbeds and Community for Information access Research Conference (NTCIR)*, pages 260–286.
- Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.
- Fandong Meng, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Interactive Attention for Neural Machine Translation. *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 2174–2185.
- Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage Embedding Models for Neural Machine Translation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 955–960.
- Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian Scientific Paper Excerpt Corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 2204–2208.
- Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-Translation for Neural Machine Translation. *Proceedings of the 26th International*

Conference on Computational Linguistics (COLING), pages 1828–1836.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum Risk Training for Neural Machine Translation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1683–1692.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural Machine Translation with Reconstruction. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 3097–3103.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling Coverage for Neural Machine Translation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 76–85.