

Neural Machine Translation with Error Correction

Kaitao Song^{1 *}, Xu Tan² and Jianfeng Lu¹

¹Nanjing University of Science and Technology

²Microsoft Research Asia

{kt.song, lujf}@njust.edu.cn, xuta@microsoft.com

Abstract

Neural machine translation (NMT) generates the next target token given as input the previous ground truth target tokens during training while the previous generated target tokens during inference, which causes discrepancy between training and inference as well as error propagation, and affects the translation accuracy. In this paper, we introduce an error correction mechanism into NMT, which corrects the error information in the previous generated tokens to better predict the next token. Specifically, we introduce two-stream self-attention from XLNet into NMT decoder, where the query stream is used to predict the next token, and meanwhile the content stream is used to correct the error information from the previous predicted tokens. We leverage scheduled sampling to simulate the prediction errors during training. Experiments on three IWSLT translation datasets and two WMT translation datasets demonstrate that our method achieves improvements over Transformer baseline and scheduled sampling. Further experimental analyses also verify the effectiveness of our proposed error correction mechanism to improve the translation quality.

1 Introduction

Neural machine translation (NMT) [Bahdanau *et al.*, 2014; Sutskever *et al.*, 2014; Vaswani *et al.*, 2017; Song *et al.*, 2018; Hassan *et al.*, 2018] have witnessed great progress due to the development of deep learning. The popular NMT models adopt an encoder-attention-decoder framework, where the decoder generates the target token based on previous tokens in an autoregressive manner. While its popularity, NMT models suffer from discrepancy between training and inference and the consequent error propagation [Bengio *et al.*, 2015; Ranzato *et al.*, 2016; Wu *et al.*, 2019]. During inference, the decoder predicts the next token given previous generated tokens as input, which is discrepant from that in training, where the previous ground-truth tokens as used as input for next token prediction. Consequently, the previous predicted tokens

may have errors, which would cause error propagation and affect the prediction of next tokens.

Previous works have tried different methods to solve the above issues, where some of them focus on simulating the data that occurs in inference for training, such as data as demonstration [Venkatraman *et al.*, 2015], scheduled sampling [Bengio *et al.*, 2015], sentence-level scheduled sampling [Zhang *et al.*, 2019], or even predict them in different directions [Wu *et al.*, 2018; Tan *et al.*, 2019]. While being effective to handle the prediction errors occurred in inference during model training, these methods still leverage the predicted tokens that could be erroneous as the conditional information to predict the next token. Forcing the model to predict correct next token given incorrect previous tokens could be particularly hard and misleading for optimization, and cannot solve the training/inference discrepancy as well as error propagation effectively.

In this paper, moving beyond scheduled sampling [Bengio *et al.*, 2015], we propose a novel method to enable the model to correct the previous predicted tokens when predicting the next token. By this way, although the decoder may have prediction errors, the model can learn the capability to build correct representations layer by layer based on the error tokens as input, which is more precise for next token prediction than directly relying on previous erroneous tokens as used in scheduled sampling.

Specifically, we introduce two-stream self-attention, which is designed for language understanding in XLNet [Yang *et al.*, 2019], into the NMT decoder to correct the errors while translation. Two-stream self-attention is originally proposed to solve the permutation language modeling, which consists of two self-attention mechanisms: the content stream is exactly the same as normal self-attention in Transformer decoder and is used to build the representations of the previous tokens, while the query stream uses the positional embedding as the inputs to decide the position of the next token to be predicted. In our work, we reinvent two-stream self-attention to support simultaneous correction and translation in NMT, where the content stream is used to correct the previous predicted tokens (correction), and the query stream is used to simultaneously predict the next token with a normal left-to-right order based on the corrected context (translation).

We conduct experiments on IWSLT 2014 German-English, Spanish-English, Hebrew-English and WMT 2014 English-

*Corresponding Author: Jianfeng Lu.

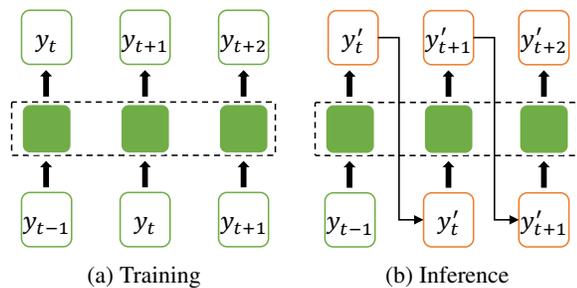


Figure 1: The discrepancy between training and inference in autoregressive sequence generation.

German and English-Romanian translation datasets to evaluate the effectiveness of our proposed error correction mechanism for NMT. Experimental results demonstrate that our method achieves improvements over Transformer baseline on all tasks. Further experimental analyses also verify the effectiveness of error correction to improve the translation accuracy.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to introduce an error correction mechanism during the translation process of NMT, with the help of the newly proposed two-stream self-attention.
- Experimental results on a variety of NMT datasets and further experimental analyses demonstrate that our method achieves improvements over Transformer baseline and scheduled sampling, verifying the effectiveness of our correction mechanism.

2 Background

In this section, we introduce the background of our work, including the standard encoder-decoder framework, exposure bias and error propagation, and two-stream self-attention mechanism.

Encoder-decoder framework. Given a sentence pair $\{x, y\} \in (\mathcal{X}, \mathcal{Y})$, the objective of an NMT model is to maximize the log-likelihood probability $P(y|x; \theta)$, where θ is the parameters of NMT model. The objective function is equivalent to a chain of the conditional probability: $P(y|x; \theta) = \prod_{t=1}^n P(y_t|y_{<t}, x; \theta)$, where n is the number of tokens in target sequence y and $y_{<t}$ means the target tokens before position t . The encoder-decoder structure [Sutskever *et al.*, 2014; Cho *et al.*, 2014] is the most common framework to solve the NMT task. It adopts an encoder to transform the source sentence x as the contextual information h and a decoder to predict the next token y_t based on the previous target tokens $y_{<t}$ and h autoregressively. Specifically, for the t -th token prediction, the decoder feeds the last token y_{t-1} as the input to predict the target token y_t . Besides, an encoder-decoder attention mechanism [Bahdanau *et al.*, 2014] is used to bridge the connection between source and target sentence.

Exposure bias and error propagation. Exposure bias [Ranzato *et al.*, 2016] is a troublesome problem in the language generation. During the training stage for language

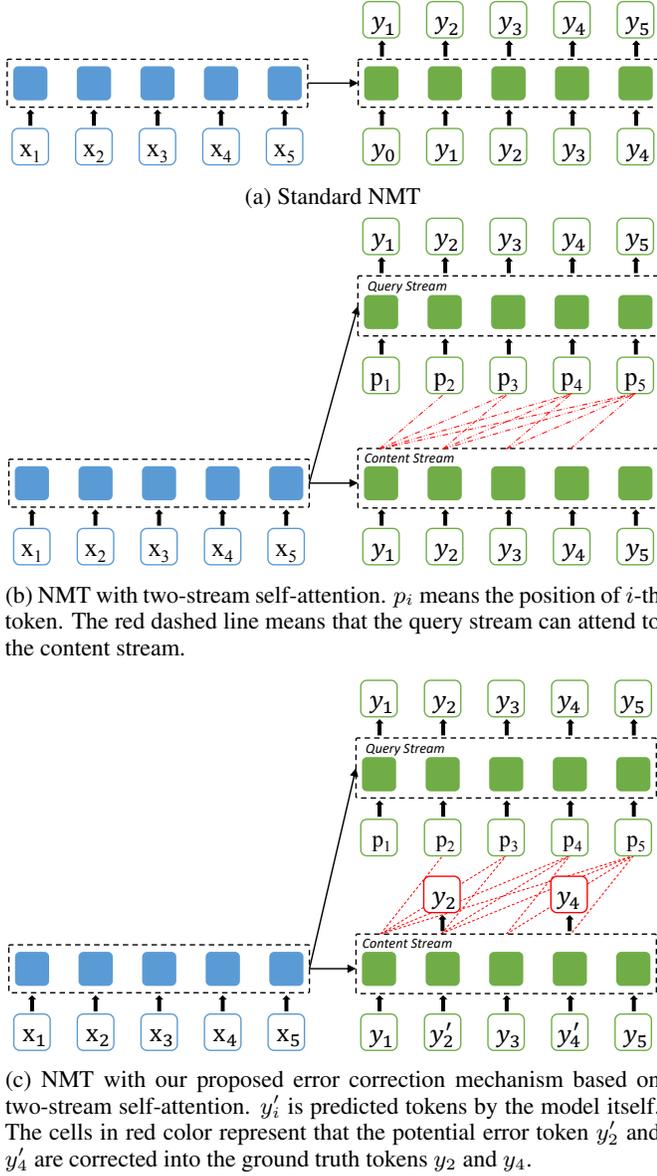
generation, it always takes ground truth tokens as the model input. However, at the test stage, the decoder depends on its previous predictions to infer next token. Figure 1 illustrates us an example of the discrepancy between training and inference in autoregressive sequence generation. Once an incorrect token is predicted, the error will accumulate continuously along the inference process. To alleviate this problem, the common solution is to replace some ground truth tokens by predicted tokens at the training time, which is named as *scheduled sampling*. However, scheduled sampling still cannot handle the exposure bias perfectly since it only attempts to predict the next ground truth according to the incorrect predicted tokens but cannot reduce the negative effect from incorrect tokens. Therefore, we propose a novel correction mechanism during translation to alleviate the error propagation.

Two-stream self-attention. XLNet is one of the famous pre-trained methods [Radford *et al.*, 2019; Devlin *et al.*, 2019; Yang *et al.*, 2019; Song *et al.*, 2019] for natural language processing. It first proposed a two-stream self-attention mechanism, which consists of a content stream and a query stream for permutation language modeling. For token y_t , it can see tokens $y_{\leq t}$ in the content stream while only see tokens $y_{<t}$ in the query stream. Beneficial from two-stream self-attention mechanism, the model can predict the next token in any position with the corresponding position embedding as the query, in order to enable permutation language modeling. Besides, Two-stream self-attention also avoids the pre-trained model to use [MASK] token into the conditioned part during the training, where the [MASK] token would bring mismatch between pre-training and fine-tuning. In this paper, we leverage the advantages of two-stream self-attention to design a novel error correction mechanism for NMT.

3 Method

Previous NMT models [Bahdanau *et al.*, 2014; Vaswani *et al.*, 2017] generate the next tokens y'_t from the probability (i.e., $y'_t \sim P(y_t|y_{<t}, x; \theta)$). If y'_t is predicted with error and taken as the next decoder input, can we force model to automatically build correct hidden representations that are close to the ground truth token y_t . In this case, the consequent token generation can build upon the previous correct representations and become more precise. A natural idea is to optimize the model to maximize the correction probability $P(y_t|y'_t, y_{<t}, x; \theta)$ simultaneously when maximizing the probability of next token prediction $P(y_{t+1}|y'_t, y_{<t}, x; \theta)$. However, the previous NMT decoder does not well support this correction mechanism. Inspired by the two-stream self-attention in XLNet [Yang *et al.*, 2019], we leverage the content stream to maximize $P(y_t|y'_t, y_{<t}, x; \theta)$ and the query stream to maximize $P(y_{t+1}|y'_t, y_{<t}, x; \theta)$, which can well meet the requirements of simultaneous correction and translation.

In order to introduce our method more clearly, in the following subsections, we first introduce the integration of two-stream self-attention into NMT model, and then introduce the error correction mechanism for NMT based on two-stream self-attention.



(b) NMT with two-stream self-attention. p_i means the position of i -th token. The red dashed line means that the query stream can attend to the content stream.

(c) NMT with our proposed error correction mechanism based on two-stream self-attention. y'_i is predicted tokens by the model itself. The cells in red color represent that the potential error token y'_2 and y'_4 are corrected into the ground truth tokens y_2 and y_4 .

Figure 2: The illustrations of standard NMT, NMT with two-stream self-attention and our proposed error correction mechanism on two-stream self-attention.

3.1 NMT with Two-Stream Self-Attention

Inspired by XLNet [Yang *et al.*, 2019], we incorporate the idea of two-stream self-attention to modify the decoding of NMT framework. Specifically, the encoder is the same as the standard NMT model, and the decoder is incorporated with two-stream self-attention where the positional embedding is taken as input in query stream for prediction while content stream is used to build context representations. Different from that in XLNet, we make two modifications: 1) we remove the permutation language modeling in XLNet since the decoder in NMT usually only uses left-to-right generation, and 2) we let the decoder to predict the whole sequence

rather than partial sentence in XLNet. Figure 2a and Figure 2b show the differences between the standard NMT and the NMT with two-stream self-attention.

We formulate the NMT with two-stream self-attention as follows. For the decoder, we feed the positions $\{p_1, \dots, p_n\}$ to the query stream to provide the position information for the next token prediction, and the sequence $\{y_1, \dots, y_n\}$ plus its positions $\{p_1, \dots, p_n\}$ to the content stream to build contextual information. For the l -th layer, we define the hidden states of query/content streams as q_t^l and c_t^l . The updates for the query and content streams are as follows:

$$q_t^{l+1} = \text{Attention}(Q = q_t^l, KV = c_{<t}^l; h, \theta_{l+1}) \quad (1)$$

$$c_t^{l+1} = \text{Attention}(Q = c_t^l, KV = c_{<t}^l; h, \theta_{l+1}), \quad (2)$$

where h represents the hidden states from encoder outputs and θ_{l+1} represents the parameters of the $l+1$ layer, Q and KV represents the query, key and value in self-attention. Both the query and content stream share the same model parameters. The states of key and value can be reused in both query and content streams. Finally, we feed the outputs of the query stream from the last layer to calculate the log-probability for the next target-token prediction. During inference, we first predict the next token with the query stream, and then update the content stream with the generated token. The order of query and content stream will not affect the predictions since the tokens in the query stream only depend on the previously generated tokens of the content streams.

3.2 Error Correction based on Two-Stream Self-Attention

Benefiting from the two-stream self-attention, we can naturally introduce an error correction mechanism on the content stream. The content stream is originally designed to build the representations of previous tokens, which is used in query stream for next token prediction. In order to correct errors, the content stream also needs to predict the correct tokens given incorrect tokens as input.

In order to simulate the prediction errors in the input of the content stream, we also leverage scheduled sampling [Bengio *et al.*, 2015] to randomly sample tokens either from the ground truth $y = \{y_1, \dots, y_n\}$ or the previously predicted tokens $y' = \{y'_1, \dots, y'_n\}$ with a certain probability as the new inputs $\tilde{y} = \{\tilde{y}_1, \dots, \tilde{y}_n\}$, where y'_t is sampled from the probability distribution $P(y_t | y_{<t}, x; \theta)$. For inputs \tilde{y}_t , it equals to y_t with a probability $p(\cdot)$ otherwise y'_t . For each token y'_t ($y'_t \neq y_t$) predicted by the query stream in step t , we force the content stream to predict its corresponding ground truth token y_t again. The loss function for the error correction mechanism (ECM) is formulated as:

$$\mathcal{L}_{\text{ECM}}(y | \tilde{y}, x; \theta) = - \sum_{t=1}^n 1(\tilde{y}_t \neq y_t) \log(P(y_t | \tilde{y}_{\leq t}, x; \theta)). \quad (3)$$

In the mechanism, the content stream can learn to gradually correct the hidden representations of error tokens toward the correct counterpart layer by layer.

The query stream is still used to predict the the next token, given a random mixture of previous predicted tokens and

ground truth tokens. The negative log-likelihood (NLL) loss is formulated as:

$$\mathcal{L}_{\text{NLL}}(y|\tilde{y}, x; \theta) = - \sum_{t=1}^n \log(P(y_t|\tilde{y}_{<t}, x; \theta)). \quad (4)$$

Finally, we combine the two loss functions as the final objective function for our method:

$$\min \mathcal{L}_{\text{NLL}}(y|\tilde{y}, x; \theta) + \lambda \cdot \mathcal{L}_{\text{ECM}}(y|\tilde{y}, x; \theta), \quad (5)$$

where λ is a hyperparameter to balance the NLL loss and ECM loss.

Figure 2c demonstrates the workflow of our proposed error correction mechanism. The difference between our error correction mechanism and the naive scheduled sampling is that once an error token is predicted in scheduled sampling, the model still learns to predict the next correct token given error tokens as context, which could confuse the the model and mislead to learn incorrect prediction patterns. However, based on our error correction mechanism, the next token prediction is built upon the representations that are corrected by the content stream, and is more precise to learn prediction patterns.

In our error correction mechanism, how to control the scheduled sampling probability $p(\cdot)$ and when to sample tokens are important factors for the training. Previous works [Bengio *et al.*, 2015] indicated that it is unsuitable to sample tokens from scratch during the training since the model is still under-fitting and the sampled tokens will be too erroneous. Inspired by OR-NMT [Zhang *et al.*, 2019], we design a similar exponential decay function for sampling probability $p(\cdot)$ but with more restrictions. The decay function is set as:

$$p(s) = \begin{cases} 1, & s \leq \alpha \\ \max(\beta, \frac{\mu}{\mu + \exp((s-\alpha)/\mu)}), & \text{otherwise} \end{cases}, \quad (6)$$

where s represents the training step, α , β and μ are hyperparameters. The hyperparameter α means the step when model starts to sample tokens, and hyperparameter β is the maximum probability for sampling.

4 Experimental Setting

In this section, we introduce the experimental settings to evaluate our proposed method, including datasets, model configuration, training and evaluation.

4.1 Datasets

We conduct experiments on three IWSLT translation datasets ($\{\text{German, Spanish, Hebrew}\} \rightarrow \text{English}$) and two WMT translation datasets ($\text{English} \rightarrow \{\text{German, Romanian}\}$) to evaluate our method. In the follow sections, we abbreviate English, German, Spanish, Hebrew, Romanian as “En”, “De”, “Es”, “He”, “Ro”.

IWSLT datasets. For IWSLT14 De \rightarrow En, it contains 160K and 7K sentence pairs in training set and valid set. We concatenate TED.tst2010, TED.tst2011, TED.tst2012,

TED.dev2010 and TED.tst2012 as the test set. For IWSLT14 Es \rightarrow En and He \rightarrow En¹, they contain 180K and 150K bilingual data for training. We choose TED.tst2013 as the valid set and TED.tst2014 as the test set. During the data preprocess, we learn a 10K byte-pair-coding (BPE) [Sennrich *et al.*, 2016] to handle the vocabulary.

WMT datasets. WMT14 En \rightarrow De and WMT16 En \rightarrow Ro translation tasks contain 4.5M and 2.8M bilingual data for training. Following previous work [Vaswani *et al.*, 2017], we concatenate newstest2012 and newstest2013 as the valid set, and choose newstest2014 as the test set for WMT14 En \rightarrow De. For WMT16 En \rightarrow Ro, we choose newsdev2016 as the valid set and newstest2016 as the test set. We learn 32K and 40K BPE codes to tokenize WMT14 En \rightarrow De and WMT16 En \rightarrow Ro dataset.

4.2 Model Configuration

We choose the state-of-the-art Transformer [Vaswani *et al.*, 2017] as the default model. For IWSLT tasks, we use 6 Transformer blocks, where attention heads, hidden size and filter size are 4, 512 and 1024. Dropout is set as 0.3 for IWSLT tasks. The parameter size is 39M. For WMT tasks, we use 6 Transformer blocks, where attention heads, hidden size and filter size are 16, 1024 and 4096. And the parameter size is 214M. Dropout is set as 0.3 and 0.2 for En \rightarrow De and En \rightarrow Ro respectively. To make a fair comparison, we also list some results by the original NMT model, without two-stream self-attention. For the decay function of sampling probability, we set a , b and μ as 30,000, 0.85 and 5,000. The λ for \mathcal{L}_{ECM} is tuned on the valid set, and a optimal choice is 1.0. To manifest the advances of our method, we also prepare some strong baselines for reference, including: Layer-wise Transformer [He *et al.*, 2018], MIXER [Ranzato *et al.*, 2016] on Transformer and Tied-Transformer [Xia *et al.*, 2019].

4.3 Training and Evaluation

During training, we use Adam [Kingma and Ba, 2014] as the default optimizer, with a linear decay of learning rate. The IWSLT tasks are trained on single NVIDIA P40 GPU for 100K steps and the WMT tasks are trained with 8 NVIDIA P40 GPUs for 300K steps, where each GPU is filled with 4096 tokens. During inference, We use beam search to decode results. The beam size and length penalty is set as 5 and 1.0 for each task except WMT14 En \rightarrow De, which use a beam size of 4 and length penalty is 0.6 by following the previous work [Vaswani *et al.*, 2017]. All of the results are reported by multi-bleu². Our code is implemented on fairseq [Ott *et al.*, 2019]³, and we will release our code under this link: <https://github.com/StillKeepTry/ECM-NMT>.

5 Results

In this section, we report our result on three IWSLT tasks and two WMT tasks. Furthermore, we also study each hyper-

¹IWSLT datasets can be download from <https://wit3.fbk.eu/archive/2014-01/texts>

²<https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl>

³<https://github.com/pytorch/fairseq>

Method	De→En	Es→En	He→En
Tied Transformer	35.10	40.51	
Layer-Wise Transformer	35.07	40.50	-
MIXER	35.30	42.30	-
Transformer Baseline	34.78	41.78	35.32
Our method	35.70	43.05	36.49

Table 1: BLEU score on IWSLT14 Translation tasks in different setting.

Method	En→De	En→Ro
Tied Transformer	28.98	34.67
Layer-wise Transformer	29.01	34.43
MIXER	28.68	34.10
Transformer Baseline	28.40	32.90
Our method	29.20	34.70

Table 2: BLEU score on WMT14 En→De and WMT16 En→Ro.

parameter used in our model, and conduct ablation study to evaluate our method.

5.1 Results on IWSLT14 Translation Tasks

The results of IWSLT14 tasks are reported in Table 1. From Table 1, we find our model with correction mechanism outperforms baseline by 0.89, 0.99 and 0.83 points and the original NMT baseline by 0.92, 1.27 and 1.17 points on De→En, Es→En and He→En respectively. Note that our baseline is strong enough which is comparable to the current advanced systems. Even within such strong baselines, our method stills achieves consistent improvements in all three tasks. These improvements also confirm the effectiveness of our method in correcting error information.

5.2 Results on WMT Translation Tasks

In order to validate the performance of our method on large-scale datasets, we also conduct experiments on WMT14 En→De and WMT16 En→Ro. The results are reported in Table 2. We found when incorporating error correction mechanism into the NMT model, it can achieve 29.20 and 34.70 BLEU score, which outperforms our baseline by 0.8 and 1.6 points in En→De and En→Ro, and is comparable to the previous works. These significant improvements on two large-scale datasets also demonstrate the effectiveness and robustness of our method in solving exposure bias. In addition, our approach is also compatible with previous works, *i.e.*, our method can achieve better performance if combined with other advanced structure.

5.3 Ablation Study

To demonstrate the necessity of each component in our method, we take a series of ablation study on our model on IWSLT14 De→En, Es→En and WMT14 En→De. The results are shown in Table 3. When disabling error correction mechanism (ECM), we find the model accuracy decreases 0.30, 0.48 and 0.50 points respectively in different

Method	De→En	Es→En	En→De
Our method	35.70	43.05	29.20
-ECM	35.40	42.55	28.72
-ECM -SS	34.81	42.16	28.48
-ECM -SS -TSSA	34.78	41.81	28.40

Table 3: Ablation study of different component in our model. The second, third and fourth column are results on IWSLT14 De→En, Es→En and WMT14 En→De. The second row is our method. The third row is equal to the second row removing error correction mechanism (ECM). The fourth row is equal to the third row removing scheduled sampling (SS). The last row is equal to the fourth row removing two-stream self-attention, *i.e.*, the standard NMT. The prefix “-” means removing this part.

tasks. When further removing scheduled sampling (SS), the model accuracy drop to 34.81, 41.16, 28.48 points in three tasks. We observe that in the large-scale dataset, the improvements for scheduled sampling are limited, while our model still achieves stable improvements in the large-scale dataset, which proves the effectiveness of the error correction mechanism. In addition, we also make a comparison between the original NMT and the NMT with two-stream self-attention (TSSA) to verify whether two-stream self-attention mechanism will contribute to model accuracy improvement. From Table 3, we find the NMT model with TSSA performs slightly better than the original NMT model in the small-scale tasks by 0.03-0.35 points and is close to the accuracy of the original NMT model in the large-scale tasks. This phenomenon also explains the improvement of our method is mainly brought by the error correction, rather than the two-stream self-attention. In summary, every component all plays an indispensable role in our model.

5.4 Study of λ for \mathcal{L}_{ECM}

To investigate the effect of λ in \mathcal{L}_{ECM} on model accuracy, we conduct a series of experiments with different λ on IWSLT14 De→En and WMT14 En→De datasets. The results are shown in Figure 3. It can be seen that the accuracy improves with the growth of λ . When $\lambda \geq 0.6$, the model accuracy increases slowly. The model achieves best accuracy at $\lambda = 0.9$ in IWSLT14 De→En and $\lambda = 1.0$ in WMT14 En→De. To

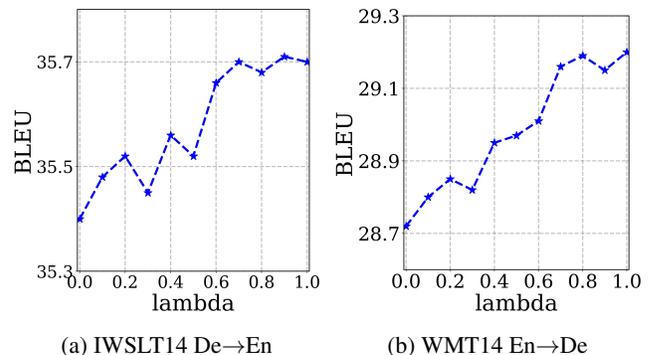


Figure 3: Results on IWSLT14 De→En and WMT14 En→De with different λ .

Method	Translation
Source (De)	in dem moment war es , als ob ein filmregisseur einen bühnenwechsel verlangt hätte .
Target (En)	at that moment , it was as if a film director called for a set change .
Baseline	it was <i>like a movie</i> director had demanded a <i>shift</i> at the moment .
Baseline + SS	at <i>the</i> moment , it was <i>like a filmmaker</i> had requested a change.
Our method	at <i>the</i> moment , it was as if a <i>movie</i> director had called for a change.

Table 4: A translation case on IWSLT14 De→En test set, generated by the baseline method, baseline with scheduled sampling and our method with error correction. The italic font means the mismatch translation.

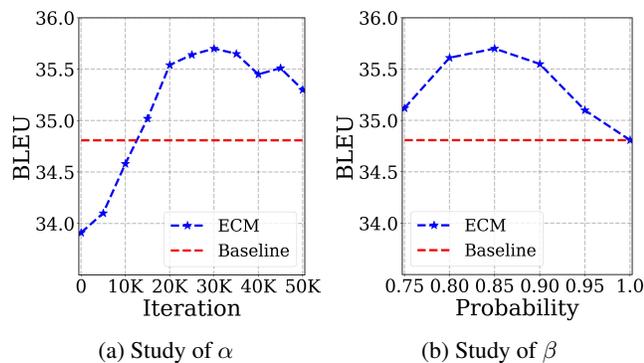


Figure 4: Results on IWSLT14 De→En with different start steps α for token sampling and maximum sampling probabilities β .

avoid the heavy cost in tuning hyperparameters for different tasks, we limit λ as 1.0 in all of final experiments.

5.5 Study of Sampling Probability $p(\cdot)$

In this subsection, we study how the hyperparameters α and β in sampling probability $p(\cdot)$ will affect the model accuracy to what extent. We conduct experiments on IWSLT14 De→En dataset to investigate when to start scheduled sampling (α) and the best choice of maximum sampling probability (β). As can be seen from Figure 4, we have the following observations:

- From Figure 4a, it can be seen that starting to sample tokens before 10K steps results in worse accuracy than the baseline, which is consistent with the previous hypothesis as mentioned in Section 3.2 that sampling tokens too early will affect the accuracy. After 10K steps, our method gradually surpasses the baseline and achieves the best accuracy at 30K steps. After 30K steps, the model accuracy drops a little. In summary, $\alpha = 30K$ is an optimal choice to start token sampling.
- From Figure 4b, the model accuracy decreases promptly, no matter when the maximum sampling probability is too big (*i.e.*, the sampled tokens is close to the ground truths) or too small (*i.e.*, the sampled tokens are close to the predicted tokens). This phenomenon is also consistent with our previous analysis in Section 3.2. Therefore, we choose $\beta = 0.85$ as the maximum sampling probability.

5.6 Cases Study

To better understand the advantages of our method in correcting error tokens, we also prepare some translation cases in IWSLT14 De→En, as shown in Table 4. It can be seen that the baseline result deviates the ground truth too much, which only predicts five correct tokens. When further adding scheduled sampling, the quality of translation can be improved, but still generates error tokens like “*like a filmmaker had requested*”, which cause the following error prediction. Finally, when using error correction mechanism, we find that the translation produced by our method is closer to the target sequence with only tiny errors. In our translation sentence, although a mismatched token “*movie*” is predicted, our model can still correct the following predicted sequence like “*called for a change*”. The quality of our translation also confirms the effectiveness of our model in correcting error information and alleviating error propagation.

6 Conclusion

In this paper, we incorporated a novel error correction mechanism into neural machine translation, which aims to solve the error propagation problem in sequence generation. Specifically, we introduce two-stream self-attention into neural machine translation, and further design a error correction mechanism based on two-stream self-attention, which is able to correct the previous predicted errors while generate the next token. Experimental results on three IWSLT tasks and two WMT tasks demonstrate our method outperforms previous methods including scheduled sampling, and alleviates the problem of error propagation effectively. In the future, we expect to apply our method on other sequence generation tasks, *e.g.*, text summarization, unsupervised neural machine translation, and incorporate our error correction mechanism into other advanced structures.

Acknowledgments

This paper was supported by The National Key Research and Development Program of China under Grant 2017YFB1300205.

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation

- by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473, September 2014.
- [Bengio *et al.*, 2015] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, pages 1171–1179. 2015.
- [Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, oct 2014.
- [Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.
- [Hassan *et al.*, 2018] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*, 2018.
- [He *et al.*, 2018] Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Layer-wise coordination between encoder and decoder for neural machine translation. In *NIPS*, pages 7944–7954. 2018.
- [Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2014.
- [Ott *et al.*, 2019] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*, June 2019.
- [Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.
- [Ranzato *et al.*, 2016] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016.
- [Sennrich *et al.*, 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, August 2016.
- [Song *et al.*, 2018] Kaitao Song, Xu Tan, Di He, Jianfeng Lu, Tao Qin, and Tie-Yan Liu. Double path networks for sequence to sequence learning. In *COLING*, pages 3064–3074, August 2018.
- [Song *et al.*, 2019] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. In *ICML*, pages 5926–5936, 2019.
- [Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. 2014.
- [Tan *et al.*, 2019] Xu Tan, Yingce Xia, Lijun Wu, and Tao Qin. Efficient bidirectional neural machine translation. *arXiv preprint arXiv:1908.09329*, 2019.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.
- [Venkatraman *et al.*, 2015] Arun Venkatraman, Martial Hebert, and J. Andrew Bagnell. Improving multi-step prediction of learned time series models. In *AAAI*, 2015.
- [Wu *et al.*, 2018] Lijun Wu, Xu Tan, Di He, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. Beyond error propagation in neural machine translation: Characteristics of language also matter. *arXiv preprint arXiv:1809.00120*, 2018.
- [Wu *et al.*, 2019] Lijun Wu, Xu Tan, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. Beyond error propagation: Language branching also affects the accuracy of sequence generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):1868–1879, 2019.
- [Xia *et al.*, 2019] Yingce Xia, Tianyu He, Xu Tan, Fei Tian, Di He, and Tao Qin. Tied transformers: Neural machine translation with shared encoder and decoder. In *AAAI*, pages 5466–5473, 2019.
- [Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NIPS*, pages 5754–5764, 2019.
- [Zhang *et al.*, 2019] Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. Bridging the gap between training and inference for neural machine translation. In *ACL*, pages 4334–4343, July 2019.