Challenge Training to Simulate Inference in Machine Translation

1st Wenjie Lu, 2nd Jie Zhou, 3nd Leiying Zhou, 4nd Gongshen Liu*, 5nd Quanhai Zhang*
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
{jonsey,sanny02,zhouleiying,lgshen,qhzhang}@sjtu.edu.cn

***Abstract*—Despite much success has been achieved, neural machine translation (NMT) suffers from exposure bias and evaluation discrepancy. To be specific, the generation inconsistency between the training and inference process further causes error accumulation and distribution disparity. Furthermore, NMT models are generally optimized on word-level cross-entropy loss function but evaluated by sentence-level metrics. This evaluation-level mismatch may mislead the promotion of translation performance. To address these two drawbacks, we propose to challenge training to gradually simulate inference. Namely, the decoder is fed with inferred words rather than ground truth words during training with a dynamic probability. To ensure accuracy and integrity, we adopt alignment and tailoring on the inferred words. Therefore, these words can leverage inferred information to help improve the training process. As for the dynamic simulation, we define a novel loss-sensitive probability that can sense the converge of training and finetune itself in turn. Experimental results on IWSLT 2016 German-English and WMT 2019 English-Chinese datasets demonstrate that our methodology can significantly improve translation quality. The approach of alignment and tailoring outperforms previous works. Meanwhile, the proposed loss-sensitive sampling is also useful for other state-of-the-art scheduled sampling methods to achieve further promotion.**

***Index Terms*—neural machine translation, exposure bias, evaluation discrepancy, schedule sampling**

I. Introduction

Neural machine translation (NMT) is an important direction in natural language processing. It can be formulated as a sequence to sequence task with a general encoder-decoder-attention architecture [1] [2]. NMT models are able to achieve promising results based on sufficient and diverse corpora.

However, NMT models suffer from two major drawbacks. First, there exists bias between training and inference situation which is called exposure bias [3]. To be specific, at training time, target-side ground truth words are fed as input to the decoder. Then outputs are collected for model to optimize. But at inference time, ground truth words no more exist. The models have to first generate a word and then feed previous generated word back as input until predicting the end-of-sentence token. This discrepancy brings about problems of error accumulation and distribution disparity between the training and inference process. Second, most NMT models are trained on maximum likelihood estimation (MLE) objective. More specifically, models calculate and accumulate cross-entropy loss between outputs and ground truth sentences word by word. A lower cross-entropy value means the predictions are closer to ground truth at word level. Model parameters are updated through backpropagation to minimize the value of loss function. However, translation performance is generally measured by sentence-level metrics such as BLEU [4], ROUGE [5], etc. This way of word-level optimization mismatches sentence-level evaluation metrics, which will limit the promotion of translation quality.

Inspired by curriculum learning [6], previous works [3] [7] have proposed to transform the training process to relieve exposure bias. Their starting point is to give training situation a challenge**—simulating inference situation**,i.e.,feeding some words other than ground truth words to decoder, just like inference process. We refer to these words as challenge words. Namely, either ground truth words or challenge words are fed to decoder with a sample probability p. The probability decreases with training process so that the model can gradually adapt to this challenging situation. One noteworthy downside is that although p decays as designed during training, currently its calculation is only related to a hyper-parameter and current index of training batch or epoch. Whether the model converges fast or slow, sample probability decays in a fixed curve only adjusted by a hyper-parameter. Another problem lies in the selection of challenge words. Reference [3] samples from ground truth and previous predicted words at word level, but lack of n-gram information cannot help to solve the second problem mentioned above. Reference [7] selects challenge words at sentence level through force decoding to improve word-level optimization. However, force decoding interferes with optimal generation process, which may destroy its integrity and accuracy.

In this paper, we present a novel approach to promote training quality with the assist of inference process. Unlike previous works, we propose a loss-sensitive sample probability to sample from ground truth and challenge words, which can be automatically fine-tuned by cross-entropy loss. This dynamic probability is more flexible during training since it can sense converge speed and make adjustment. Moreover, to ensure the integrity and accuracy of the generated challenge words, we select them at sentence level by two steps. First, we adopt beam search to infer candidate sentences and choose the sentence with highest n-gram translation quality. Then, an alignment module tailors the sentence to a desirable array for

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

sampling without decreasing the quality of generation. Finally, at training process, either ground truth words or challenge words are fed at every decoding step with loss-sensitive sampling. The model is trained on MLE objective for simplicity while relieving the problem of evaluation discrepancy by transmitting n-gram evaluation information to training.

The major contributions of this paper are summarized as follows:

- We present a method to challenge training to simulate inference, aiming at alleviating exposure bias and evaluation discrepancy. At training process, either ground truth or challenge words are fed to decoder with Bernoulli distribution. The challenge words are generated at sentence level by inference, alignment and tailoring, which can capture n-gram inferred information while maintaining the accuracy of generation.
- We propose a novel loss-sensitive sample probability for sampling from ground truth and challenge words. To make the sample probability more flexible and suitable for different training situations, its calculation considers cross-entropy loss as well as the number of trained epochs. It can sense the current learning state of model and reflect in the level of challenge.
- We demonstrate the effectiveness of our approach on IWSLT 2016 German-English and WMT 2019 English-Chinese datasets, and achieve significant improvements. Moreover, adding our approach of loss-sensitive sampling to other state-of-the-art scheduled sampling methods can help achieve further promotion.

II. Related Work

A. Exposure bias.

In recent years, the problem of exposure bias has received great attention. The direct way to alleviate exposure bias is to utilize its own predictions in training. Reference [3] first posed this problem and proposed a scheduled sampling strategy based on an algorithm called Data As Demonstrator (DAD) [8]. At every decoding step, a probability p is used to decide whether to sample from ground truth or the previous word predicted by the model itself. They also compared effects of three different decay curves of p, including linear decay, exponential decay and inverse sigmoid decay. Inspired by their method, reference [7] came up with sampling from ground truth and inference sentences word by word with inverse sigmoid decay curve so that its n-gram matching nature can alleviate evaluation discrepancy. To ensure these sentence pairs have same number of words, they adopted force decoding to 'cut down' or 'force' the generation of inference sentences. We refer to their methods and further propose a novel sentence-level challenge words generation approach through alignment and tailoring instead of force decoding. Furthermore, we modify the decay curve of p to get a loss-sensitive sample probability.

B. Evaluation discrepancy.

As for the discrepancy between word-level MLE objective and sentence-level evaluation metrics, some researches utilize techniques like generative adversarial network (GAN) [9] or reinforcement learning (RL) [10]. Borrowed idea from DAD [8] and beam search [11] [12], reference [13] proposed Mixed Incremental Cross-Entropy Reinforce (MIXER) to directly optimized model parameters with respect to the metric used at inference time. Further, reference [14] presented minimum risk training (MRT) to minimize the expected loss (i.e., risk) on the training data. Reference [15] introduced beam-search optimization schedule for model to learn global sequence scores. Moreover, reference [16] proposed RankGAN which can analyze and rank sentences by giving a reference group, and thus achieve high-quality language descriptions. Reference [17] presented CoT which can work without the necessity of pre-training via MLE.

III. Methodology

The main schematic of our proposed methodology is shown in Fig. 1. The right part in the figure indicates the process of generating and selecting candidate sentences, and the left part shows a rough process of sentence alignment, tailoring and sampling.

Specifically, for each sentence pair (X, Y) to be trained, we first perform the process of inference and use beam search to predict the translation of X with beam size k. Then we choose the sentence Y^* in k candidates which scores highest BLEU value with ground truth Y to ensure its translation quality. After that, we conduct alignment and tailoring on Y^* according to the ground truth sentence Y, generating the desirable challenge sentence (composed of challenge words) $Y^{'}$. The ground truth and challenge sentence are sampled with Bernoulli distribution at probability p. Finally, the model trains with sentence X and the sampled words as new parallel data. After training an epoch, the training loss is fed back to sample module and it recalculates the sample probability p. Thus, training and inference can interact with each other and get promotion.

In the following, we will first describe the NMT model in Section III-A, and then introduce the generation of challenge sentence in Section III-B. The method of sentence alignment and tailoring is presented in details in Section III-C. Finally, Section III-D explains how to define loss-sensitive sample probability.

A. Model Overview

We utilize a common RNN attention model [2] as baseline to demonstrate our approach. Given the source sentence $X = (x_1, x_2, ..., x_{T_x})$ and the target sentence $Y = (y_1, y_1, ..., y_{T_y})$, the model can deal with many sequence to sequence problems. For the input sentence X, the encoder first converts each word into its own word vector $w_t \subseteq R^K, t = 1, 2, ..., T_x$. After obtaining word embeddings, RNN encodes the source sentence as follows:

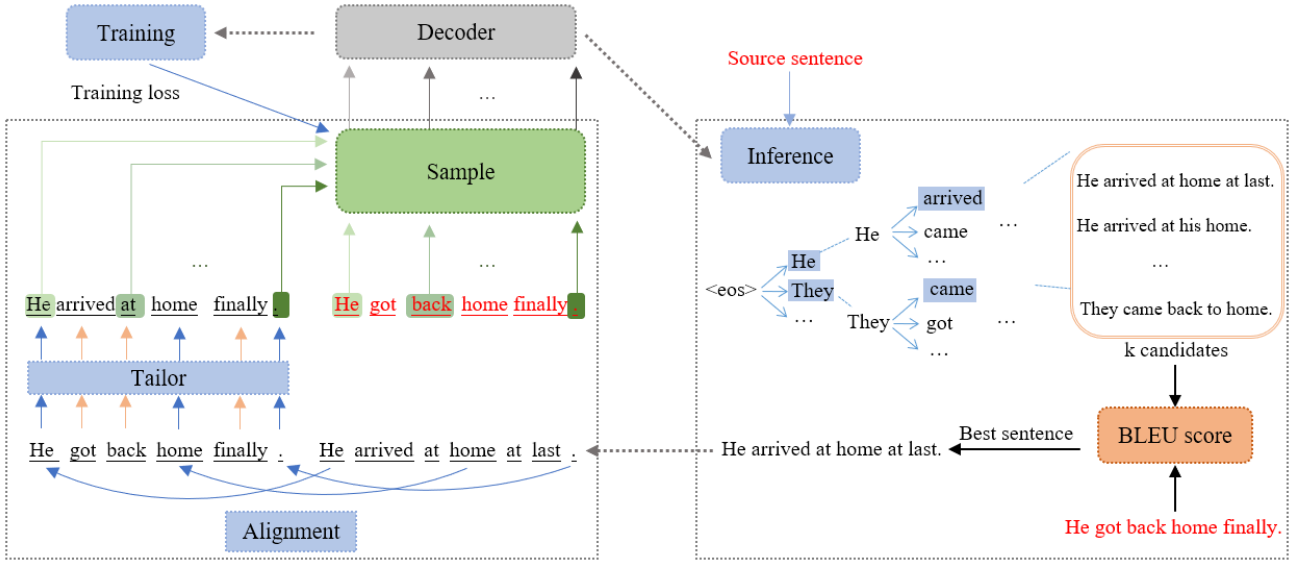


Fig. 1. A schematic of our approach. The sentences written in red indicate the ground truth source and target sentence. The dashed arrows show the overall iteration process between the training and inference process. Benefit from their interaction, training and inference can achieve mutual promotion.

$$h_t = \phi(h_{t-1}, w_t) \tag{1}$$

where h_0 is an initial vector, ϕ is a nonlinear function of hidden layers. Then context vector $c_i, i = 1, 2, ..., T_y$ is calculated by:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} \cdot h_j \tag{2}$$

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{3}$$

where e_{ij} is an alignment model used to evaluate the match level between the j-th input word and the i-th output word.

When the decoder receives the context c_t, it calculates the hidden layer vector s_t by:

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \tag{4}$$

where s_0 is an initial vector, f is a nonlinear function of hidden layers, y_{t-1} is the historical output at time $t-1$ in inference and ground truth word in training, and y_0 is the end flag of source sentence X.

According to the hidden layer state s_t, the probability of inferring the word y_t can be computed by:

$$P(y_t) = softmax(W_o \cdot p(y_t \mid y_1, ..., y_{t-1}, x)) \tag{5}$$

$$p(y_t \mid y_1, ..., y_{t-1}, x) = g(y_{t-1}, s_t, c_t) \tag{6}$$

where g is a nonlinear function and W_o is a mapping matrix.

Finally, supposing that there are sequence pairs $(X_i, Y_i), i = 1, 2, ..., N$ in the parallel corpus, the objective of training a NMT model is to maximize the likelihood as follows:

$$L(\theta) = \sum_{i=1}^{N} log \;\; p(Y_i \mid X_i, \theta) \tag{7}$$

B. Challenge Sentence Generation

To infer the best translation of a given source sentence, the most common used algorithm is beam search. The algorithm has a parameter called beam size k which means reserving k candidate translations. Supposing that the vocabulary size is V, at each decoding step t, model stitches words in vocabulary to k existing partial translation so that $k \times V$ combinations are generated. Model calculates their probability and choose top k translations as new candidates.

Specifically, to punish very short translations, beam search is maximizing the probability defined as followings:

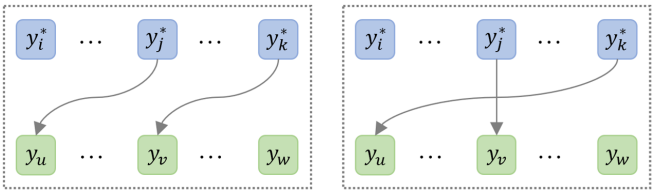
$$prob = \arg\max_{y} \frac{1}{T_y} \sum_{t=1}^{T_y} logP(y_t \mid x, y_1, ..., y_{t-1}) \tag{8}$$

where T_y is the length of output sentence.

It is interesting to note that if k is set to 1, this essentially becomes the greedy search algorithm which is generally used in training process. The model directly calculates cross-entropy loss between the only one candidate sentence and ground truth sentence. Conversely, a larger k can theoretically help achieve better translation results while consuming more memory and resources for reserving candidate translations. Therefore, we utilize a proper k to balance pros and cons. After beam search, we choose the sentence which scores highest on BLEU with ground truth sentence from k candidates. Similar to reference [7], we adopt the Gumbel-Max technique [18] [19] for generating more robust outputs. To be specific, the Gumbel noise is defined as follows:

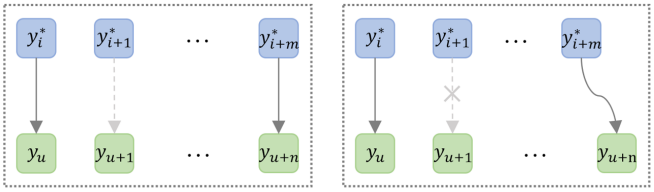
$$G = -log(-logU) \tag{9}$$

where $U \sim Unif[0, 1]$.



(a) An ideal case. (b) A special case.

Fig. 2. Two circumstances of sentence alignment, where $(y_i^*, ..., y_j^*, ..., y_k^*)$ are part of Y^* and $(y_u, ..., y_v, ..., y_w)$ are part of Y. Both of them are in proper order.



(a) An ideal case. (b) A special case.

Fig. 3. Two circumstances of sentence tailoring, where $(y_i^*, y_{i+1}^*, ..., y_{i+m}^*)$ are part of Y^* and $(y_u, y_{u+1}, ..., y_{u+n})$ are part of Y. Both of them are in proper order. In 3(a) $m = n$, while in 3(b) $m \neq n$.

Then equation (5) is modified to:

$$P(y_t) = softmax(\frac{W_o \cdot p(y_t \mid y_1, ..., y_{t-1}, x) + G}{\tau}) \quad (10)$$

where τ is a temperature parameter controlling the generated distribution.

We consider that choosing beam search translation for selecting challenge sentences have two benefits. First, beam search can generate translations better than greedy search since it can avoid local optimum and calculate probabilities globally. Second, BLEU measures n-gram matching precision of the inference and reference sentence. Therefore, selecting the best translation from beam search results can ensure its quality and n-gram accuracy. Indirectly, sentence-level information are transmitted to training process to alleviate evaluation discrepancy.

C. Sentence Alignment and Tailoring

Beam search translation provides the best sentence $Y^* = (y_1^*, y_2^*, \ldots, y_{T_{Y^*}}^*)$ which scores highest on BLEU with corresponding ground truth sentence. However, how many tokens it will generate is uncertain. This is against the requirement of sampling from ground truth and challenge sentences word by word. Therefore, we align and tailor Y^* to generate a desired array according to the ground truth sentence $Y = (y_1, y_2, ..., y_{T_Y})$ as described below.

1) Sentence Alignment: First of all, we import a mask vector $m = (m_1, m_2, ..., m_{T_Y})$ to indicate which words of Y can be aligned. m is initialized as $(0, 0, ..., 0)$ which means every word in Y is free for alignment. Then we iterate over Y^* to find words that align to masked Y. Once a word $Y_i^*(1 \leq i \leq T_{Y^*})$ is found to be aligned to $Y_j(1 \leq j \leq T_Y)$, we revise $m_1, ..., m_j$ to 1, so words before position j do not participate in the alignment process any more.

In other words, Fig. 2 shows two circumstances of sentence alignment. Fig. 2(a) is an ideal case where every alignment word pairs are in order. However, special cases may happen as shown in Fig. 2(b). We will break the alignment between y_k^* and y_u, and find if there are words aligned to y_k^* starting from y_v.

After then, we create a new array $Y^{'} = (y_1^{'}, y_2^{'}, ..., y_{T_{Y^{'}}}^{'})$ to fill in these aligned words at their specific position. If y_j^* is aligned to y_v, then position v is filled in with y_j^*.

2) Sentence Tailoring: After filling in aligned words in Y^*, we pay attention to these unaligned fragments between two aligned words. Whether they are ordinary fragments or fragments at head and tail, they are facing two circumstance as demonstrated in Fig. 3. We assume that y_i^* is aligned to y_u and y_{i+m}^* is aligned to y_{u+n}. For an ideal circumstance in Fig. 3(a), fragment $(y_{i+1}^*, ..., y_{i+m-1}^*)$ has same number of words as fragment $(y_{u+1}, ..., y_{u+n-1})$. It seems two fragments can be aligned word by word although actually these words are different. We directly fill in the fragment of Y^* to $Y^{'}$ at right position. Another circumstance is showed in Fig. 3(b) where two fragments have different length. In consider of preserving accuracy, we do not refer to fragment Y^*, but fill in fragment of Y to $Y^{'}$.

As the model converges, Y^* will be similar to Y in overall grammatical structure and the number of aligned words will greatly increase. Moreover, the occurrence of special cases in Fig. 3(b) will also decrease. Those unaligned fragments in Fig. 3(a) can help transmit n-gram inference information to the training process. After alignment and tailoring, whether the inferred sentence Y^* is longer or shorter than ground truth sentence Y, we can obtain challenge sentence $Y^{'}$ which has same words as Y.

D. Loss-sensitive Sampling

To challenge the training process to simulate inference when predicting token y_t, we propose to sample from ground truth word y_{t-1} and challenge word $y_{t-1}^{'}$. Inspired by reference [3], we use Bernoulli distribution for sampling with probability p. Assuming w_t is the input at each decoding step t, then $Pr(w_t = y_{t-1}) = p$ and $Pr(w_t = y_{t-1}^{'}) = 1 - p$. We hope the probability p to decay from 1 to 0, so that the training process can gradually learned to deal with simulated inference situation.

Borrowing idea from the decay schedule in learning rate, sample probability can be defined as an inverse sigmoid curve with variable training epochs. Furthermore, a loss function intuitively reflects how well the model is trained. If predictions are pretty good, the loss function will output a lower value. Conversely, if model fails to predict correct words, it will output a higher value. This feature can give a good feedback on fine-tuning probability p. If p decreases faster than loss, it means that model may be exposed to inference scenario too early and hard to correct these mistakes. Conversely, if p decreases more slowly than loss, we can conclude that model

TABLE I
RESULTS OF THE PROPOSED METHOD IN COMPARISON TO BASELINE SYSTEMS (BLEU).

Systems	DE-EN					EN-ZH			
	testset10	testset11	testset12	testset14	average	newstest17	newstest18	newstest19	average
Transformer	25.17	30.03	26.20	24.24	26.41	26.37	25.09	25.76	25.74
Evolved Transformer	26.33	31.45	27.28	25.36	27.61	27.84	25.98	27.25	27.02
DTMT	26.51	31.66	27.64	26.02	27.96	28.07	26.10	27.34	27.17
RNNsearch	24.46	28.06	24.92	22.94	25.10	24.92	24.17	24.20	24.63
+ SS-NMT	25.73	29.48	26.34	23.66	26.30	25.53	24.75	24.93	25.07
+ OR-NMT	26.89	29.91	26.77	24.66	27.06	27.61	25.74	26.42	26.59
+ AT-NMT	27.20	31.78	27.98	25.78	28.19	28.10	25.93	27.23	27.09
+ AT-NMT + lss	**27.68**	**32.07**	**28.18**	**26.16**	**28.52**	**28.42**	**26.21**	**27.45**	**27.36**

Note that + lss refers to the abbreviation of combining loss-sensitive sampling. Overall best results are **in bold**. All systems in comparison are trained on two public corpora using their source codes, thus the BLEU results are different from those reported in their papers.

is temporarily strong enough to handle difficulties given by inference and needs further challenge.

Therefore, we define sample probability as follows:

$$p = \frac{k}{k + exp(\frac{e}{k})} \cdot \sigma(L) \tag{11}$$

where k is a hyper-parameter, e is the current index of epoch, L is the average loss function value of epoch e, and σ is a non-linear function. Considering that feeding tailored challenge words will to some extent reduce the difficulty of inference, we hope the sample probability be lower so as to increase the level of challenge. We choose tanh function so that $0 < \sigma(L) < 1$.

IV. EXPERIMENTS

A. Experimental Setup

We conduct our experiments comparable with previous work by using the following two datasets:

a) German-English: the German-English dataset is chosen from IWSLT 2016 [20]. We use official testset2013 as validation set. The training and validation data consists of 196,884 and 992 sentences respectively. As for evaluation, we use the testset dataset from 2010 to 2014 and tokenized BLEU scores as computed by the multi-bleu.perl script[1].

b) English-Chinese: the English-Chinese dataset is chosen from the casia2015 parallel corpus in WMT 2019 shared task. It consists of approximately 1.05M sentences. We use official newsdev2017 as validation set and evaluate on the newstest dataset from 2017 to 2019.

For all training data, we perform tokenization and truecasing using standard Moses tools. For Chinese corpora, we use jieba[2] for segmentation. Then, we employ byte pair encoding (BPE) [21] with 50,000 operations to alleviate Out-of-Vocabulary problem. To accelerate training and save cost, we discard sentences with more than 50 tokens. The dimension of word embeddings is set to 512.

The training of the proposed system is similar to general RNN models with the cross-entropy loss function and a batch size of 60. We use Adam [22] optimizer to tune the parameters. Besides, we use dropout regularization with a drop probability 0.5. During decoding, the beam size is set to 3. The hyper-parameter of sample probability k and temperature τ are set to 12 and 0.5 respectively.

B. Baseline Systems

We compare our method with existing common NMT systems including Transformer [23], Evolved Transformer [24] and DTMT [25]. Moreover, there are some previous state-of-the-art schedule sampling works. These baseline systems are included as follows:

a) RNNsearch: a vanilla attention-based recurrent neural network which consists of 2-layer bidirectional GRU units [26]. The dimension of hidden layer is 512.

b) SS-NMT: a word-level scheduled sampling method [3] which utilizes inverse sigmoid decay schedule to sample from challenge word and ground truth word. Challenge word is chosen from previous predicted word.

c) OR-NMT: a sentence-level sampling method [7] which utilizes inverse sigmoid decay schedule to sample from challenge sentence and ground truth sentence. Challenge sentence is generated by beam search and force decoding.

d) AT-NMT: our proposed sentence-level method which introduces loss-sensitive sampling schedule to sample from challenge sentence and ground truth sentence. Challenge sentence is achieved by alignment and tailoring.

C. Main Results

TABLE I reports the results of the proposed system in comparison to other NMT systems. As it can be seen, our full system (AT-NMT + lss) obtains the best published results on all testsets.

On German-English dataset, our full system can outperform RNNsearch by +3.42 BLEU averagely. On English-Chinese dataset, our full system can have an improvement of +2.73 BLEU on three testsets.

To validate the effectiveness of our sentence alignment and tailoring method, we use the original reverse sigmoid sampling schedule for comparison. As shown in the experimental results,

[1]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

[2]https://github.com/fxsjy/jieba

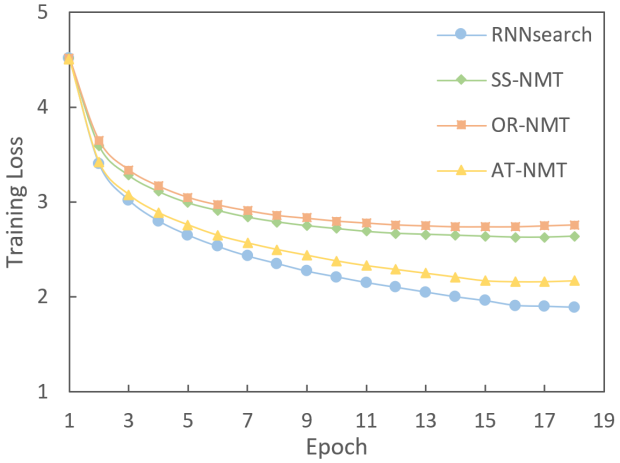


Fig. 4. The training loss curves of four baseline systems (RNNsearh, SS-NMT, OR-NMT and AT-NMT) on the IWSLT 2016 German-English translation task.

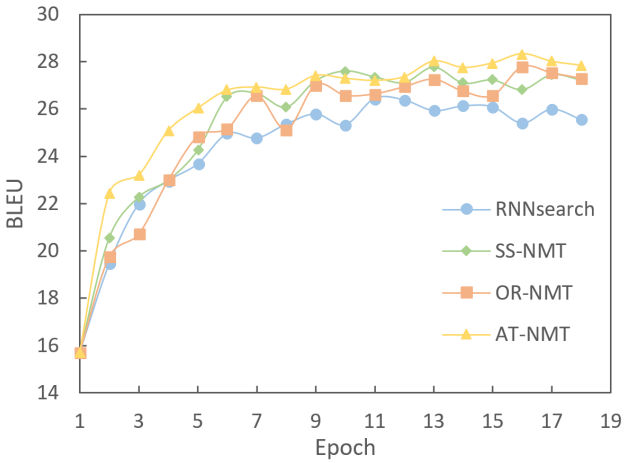


Fig. 5. Trends of BLEU scores of four baseline systems (RNNsearh, SS-NMT, OR-NMT and AT-NMT) on the validation set on the IWSLT 2016 German-English translation task.

AT-NMT can outperfrom previous works SS-NMT and OR-NMT by +1.13 $\sim$ +1.89 BLEU on German-English dataset and +0.5 $\sim$ +2.02 BLEU on English-Chinese dataset. We will display and analyse the effect of sentence alignment and tailoring in detail in Section IV-D.

While previous methods can improve translation quality by word or sentence level sampling, their sampling schedule lacks of flexibility in decaying. To solve this drawback, we propose a novel loss-sensitive sample probability which can sense the speed of converge and make adjustment on sample probability. As shown in TABLE I, AT-NMT combining loss-sensitive sampling ($+lss$) can achieve best translation performance compared to other NMT systems. We will discuss the effect of loss-sensitive sampling from two aspects in Section IV-E.

On the one hand, the idea of loss-sensitive sampling can be added to previous schedule sampling methods. Since our approach of sentence alignment and tailoring is different from previous generation method of challenge words, whether the loss-sensitive sampling is beneficial to these works are under verification. Therefore, we conduct experiments on combining our loss-sensitive sampling schedule with previous methods.

On the other hand, we consider to validate the universality of loss-sensitive sampling since it is influenced by the cross-entropy loss function value in finetuning probability p. Generally, the scale and quality of parallel corpora can have impact on the value of training loss. It is worth noting that our German-English dataset is relatively smaller, while the English-Chinese dataset is almost 5 times larger. Hence, we experiment on these two datasets to validate the effectiveness of our sampling schedule.

D. Effect of Alignment and Tailoring

To explore the effect of sentence alignment and tailoring, we make comparisons to RNNsearch, SS-NMT and OR-NMT on German-English dataset under the same conditions. We adopt the original inverse sigmoid curve instead of loss-sensitive sampling as sample probability. Fig. 4 gives the cross-entropy loss curves of RNNsearch, SS-NMT, OR-NMT and AT-NMT during training. As the training epoch increases, RNNsearch continues to decrease at a lowest value, which indicates that RNNsearch may fall in overfitting problem. Due to different generation methods of challenge words, SS-NMT, OR-NMT and AT-NMT gradually converge to a certain training loss value. Among them, the training loss of our proposed AT-NMT is much lower than other two systems. We can conclude that our tailored sentences are more reasonable and accurate, thus easier for model to correct mistakes imported by the inference process.

It is interesting to find out that the training loss of AT-NMT is lower than SS-NMT. SS-NMT selects challenge words from the model's previous predicted words which is at word level, while OR-NMT integrates force decoding to beam search which is at sentence level. Therefore, it is normal to see the training loss of OR-NMT be higher than SS-NMT since OR-NMT can import sentence-level information as compensation. However, our proposed sentence-level AT-NMT can reduce the growth of training loss. This manifests the effectiveness and high accuracy of sentence alignment and tailoring.

Moveover, Fig. 5 gives the BLEU score curves of four systems under same settings. It can be seen that RNNsearch encounters the problem of overfitting as mentioned above. As for AT-NMT, although the method of alignment and tailoring reduces the difficulty of challenge compared to other systems, it can achieve better BLEU scores on validation set. Compared to OR-NMT which uses force decoding to generate challenge sentences, our method of alignment and tailoring does preserve integrity and accuracy.

E. Effect of Loss-sensitive Sampling

Aiming at designing a more flexible sample probability, we propose the loss-sensitive sampling schedule which can sense the converge state of training and reflect on the level of challenge. We conduct experiment on German-English and

TABLE II
BLEU SCORES ON GERMAN-ENGLISH DATASET.

Systems	tst10	tst11	tst12	tst14	avg
SS-NMT	25.73	29.48	26.34	23.66	26.30
SS-NMT + lss	26.46	30.14	26.60	24.31	26.88
OR-NMT	26.89	29.91	26.77	24.66	27.06
OR-NMT + lss	27.37	30.72	27.54	25.20	27.71

TABLE III
BLEU SCORES ON ENGLISH-CHINESE DATASET.

Systems	test17	test18	test19	avg
SS-NMT	25.53	24.75	24.93	25.07
SS-NMT + lss	25.89	25.12	25.43	25.48
OR-NMT	27.61	25.74	26.42	26.59
OR-NMT + lss	28.03	26.10	26.66	26.93

English-Chinese datasets to validate the effectiveness of loss-sensitive sampling and analyse in two aspects.

We apply loss-sensitive sampling to previous schedule sampling methods SS-NMT and OR-NMT. The experimental results are listed in TABLE II and TABLE III. From the overall results, it can be seen that adding the idea of loss-sensitive sampling to these two methods can help achieve higher BLEU scores. To be specific, SS-NMT + lss can get a promotion of +0.41 $\sim$ +0.58 BLEU averagely over SS-NMT on German-English and English-Chinese datasets. OR-NMT + lss can outperform OR-NMT by +0.34 $\sim$ +0.65 BLEU score on two datasets averagely.

To deeply explore the promotion of loss-sensitive sampling on other state-of-the-art works, we observe their decay curves of sample probability during training. As shown in Fig. 6, the actual sample probabilities of SS-NMT, OR-NMT and AT-NMT on German-English dataset differ from the original sample probability. Above all, they are all lower than the original one which indicates the model tends to take challenge words as context more frequently than before. From the perspective of simulating, we make it harder for model to handle inference sentences and correct mistakes, while the experimental results show promotion on translation quality.

Another point of focus lies in the decay range of different systems. The original sample probability is calculated by a hyper-parameter and the index of current epoch. Therefore, once the value of hyper-parameter is determined, sample probability is decayed in a fixed curve. With the assist of loss-sensitive sampling, three baseline systems have specific probability decaying trends. The training loss of AT-NMT + lss is lower than SS-NMT + lss and OR-NMT + lss, so its calculated sample probability is also lower than others. On the one hand, the difficulty of challenge is reflected on the curve. Loss-sensitive sampling helps to adjust a proper sample probability for different training scenes. On the other hand, a lower sample probability means the model is mathematically exposed to inference situation more frequently. In this sense, AT-NMT + lss is more capable of simulating the inference process and achieves higher translation performance.

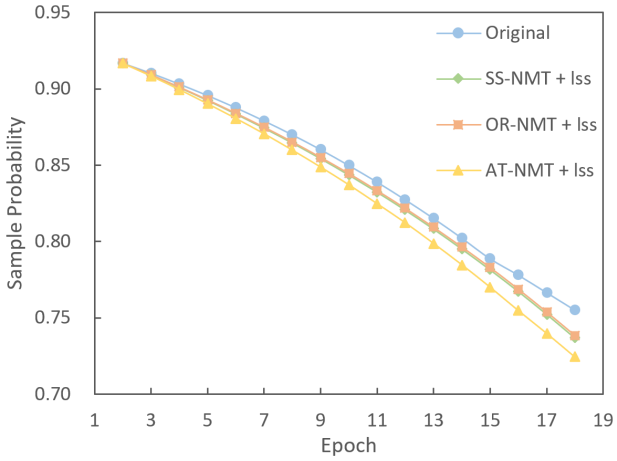


Fig. 6. Trends of sample probability on the training set on the IWSLT 2016 German-English translation task. The blue curve (original) represents previous inverse sigmoid sampling schedule. The curves of SS-NMT, OR-NMT and AT-NMT are combined with loss-sensitive sampling.

The last question we want to explore is the influence of different initial cross-entropy loss on loss-sensitive sample probability. The different scale and quality of various parallel corpus result in relatively higher or lower loss value. We observe the performance on the smaller German-English and the larger Chinese-English datasets. Generally, training loss on a larger dataset is lower than on a smaller dataset. Experiments on the two datasets also confirm this. As shown in TABLE I, loss-sensitive sampling helps to promote translation quality on both datasets. Experimental results in TABLE II and TABLE III also comfirm the effectiveness and universality of loss-sensitive sampling.

V. CONCLUSION

In this paper, we propose to challenge training to simulate inference in NMT so as to alleviate the problem of exposure bias and evaluation discrepancy. We feed challenge words rather than ground truth words as context to decoder with a sample probability in training process. The challenge words are generated in sentence level aiming to capture n-gram information. To ensure accuracy and integrity, we adopt alignment and tailoring method for inferred sentences. Moreover, we design a novel loss-sensitive sampling schedule to provide more flexible and dynamic sample probability. Experimental results show that our proposed method can achieve significant improvement on BLEU scores compared to previous works. Furthermore, our idea of loss-sensitive sampling is also helpful in promoting previous works.

ACKNOWLEDGMENT

This research work has been funded by the National Natural Science Foundation of China (Grant No.61772337), the National Key Research and Development Program of China NO. 2018YFC0830803.

REFERENCES

[1] M. L. Forcada and R. P. Ñeco, "Recursive hetero-associative memories for translation," in *International Work-Conference on Artificial Neural Networks*. Springer, 1997, pp. 453–462.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[3] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[4] D. Shterionov, P. Nagle, L. Casanellas, R. Superbo, and T. O'Dowd, "Empirical evaluation of nmt and pbsmt quality for large-scale translation production," in *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT): User Track*, 2017.

[5] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[6] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 41–48.

[7] W. Zhang, Y. Feng, F. Meng, D. You, and Q. Liu, "Bridging the gap between training and inference for neural machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4334–4343.

[8] A. Venkatraman, M. Hebert, and J. A. Bagnell, "Improving multi-step prediction of learned time series models," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 3024–3030.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 2, no. 4.

[11] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.

[12] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 379–389.

[13] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *arXiv preprint arXiv:1511.06732*, 2015.

[14] S. Shen, Y. Cheng, Z. He, W. He, H. Wu, M. Sun, and Y. Liu, "Minimum risk training for neural machine translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1683–1692.

[15] S. Wiseman and A. M. Rush, "Sequence-to-sequence learning as beam-search optimization," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1296–1306.

[16] K. Lin, D. Li, X. He, Z. Zhang, and M.-T. Sun, "Adversarial ranking for language generation," in *Advances in Neural Information Processing Systems*, 2017, pp. 3155–3165.

[17] S. Lu, L. Yu, S. Feng, Y. Zhu, and W. Zhang, "Cot: Cooperative training for generative modeling of discrete data," in *International Conference on Machine Learning*, 2019, pp. 4164–4172.

[18] E. J. Gumbel, "Statistical theory of extreme values and some practical applications," *NBS Applied Mathematics Series*, vol. 33, 1954.

[19] C. Maddison, D. Tarlow, and T. Minka, "A* sampling," *Advances in neural information processing systems*, 10 2014.

[20] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[21] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1715–1725.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[24] D. So, Q. Le, and C. Liang, "The evolved transformer," in *International Conference on Machine Learning*, 2019, pp. 5877–5886.

[25] F. Meng and J. Zhang, "Dtmt: A novel deep transition architecture for neural machine translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 224–231.

[26] K. Cho, B. v. M. C. Gulcehre, D. Bahdanau, F. B. H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2014, pp. 1724–1734.