Token-level Adaptive Training for Neural Machine Translation

Shuhao Gu[1,2], Jinchao Zhang[3], Fandong Meng[3], Yang Feng[1,2][*],
Wanying Xie[1,4], Jie Zhou[3], Dong Yu[4]
[1] Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] University of Chinese Academy of Sciences
[3] Pattern Recognition Center, WeChat AI, Tencent Inc, China
[4] Beijing Language and Culture University, China
{gushuhao19b,fengyang}@ict.ac.cn
{dayerzhang,fandongmeng,withtomzhou}@tencent.com
xiewanying07@gmail.com, yudong@blcu.edu.cn

Abstract

There exists a token imbalance phenomenon in natural language as different tokens appear with different frequencies, which leads to different learning difficulties for tokens in Neural Machine Translation (NMT). The vanilla NMT model usually adopts trivial equal-weighted objectives for target tokens with different frequencies and tends to generate more high-frequency tokens and less low-frequency tokens compared with the golden token distribution. However, low-frequency tokens may carry critical semantic information that will affect the translation quality once they are neglected. In this paper, we explored target token-level adaptive objectives based on token frequencies to assign appropriate weights for each target token during training. We aimed that those meaningful but relatively low-frequency words could be assigned with larger weights in objectives to encourage the model to pay more attention to these tokens. Our method yields consistent improvements in translation quality on ZH-EN, EN-RO, and EN-DE translation tasks, especially on sentences that contain more low-frequency tokens where we can get 1.68, 1.02, and 0.52 BLEU increases compared with baseline, respectively. Further analyses show that our method can also improve the lexical diversity of translation.

1 Introduction

Neural machine translation (NMT) systems (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) are data driven models, which highly depend on the training corpus. NMT models have a tendency towards over-fitting to frequent observations (e.g. words, word co-occurrences) while neglecting those low-frequency observations. Unfortunately, there exists a token imbalance phenomenon in natural languages as different tokens appear with different frequencies, which roughly obey the Zipf's Law (Zipf, 1949). Table 1 shows that there is a serious imbalance between high-frequency tokens and low-frequency tokens. NMT models rarely have the opportunity to learn and generate those ground-truth low-frequency tokens in the training process.

Token Order (Descending)	Average Frequency	Reference	Vanilla NMT
$[0, 10\%)$	$10,857$	81.75%	87.26%
$[10\%, 30\%)$	516	11.40%	9.06%
$[30\%, 50\%)$	133	3.43%	2.21%
$[50\%, 70\%)$	60	1.95%	0.99%
$[70\%, 100\%]$	25	1.47%	0.48%

Table 1: The average frequency on the NIST training set and proportion of tokens with different frequencies in reference and the translation of the vanilla NMT model (a Transformer model) on the NIST test sets. All the target tokens (BPE sub-words with 30K merge operations) of the training set are ranked by their frequencies in descending order. The 'Token Order' column represents the frequency interval ($[10\%, 30\%)$ means the frequency of token is between top 10% and 30%). The 'Average Frequency' column represents the average frequencies of the tokens in each interval, which show the token imbalance phenomenon in natural language. The last two columns show the vanilla NMT model tends to generate more high-frequency tokens and less low-frequency tokens than reference.

[*] Corresponding author: Yang Feng.

Joint work with Pattern Recognition Center, WeChat AI, Tencent Inc, China.

Reproducible code: https://github.com/ictnlp/TLAT-NMT.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1035–1046,
November 16–20, 2020. ©2020 Association for Computational Linguistics

Some work tries to improve the rare word translation by maintaining phrase tables or back-off vocabulary (Luong et al., 2015; Jean et al., 2015; Li et al., 2016; Pham et al., 2018) or adding extra components (Gülçehre et al., 2016; Zhao et al., 2018), which bring in extra training complexity and computing expense. Some NMT techniques which are based on smaller translation granularity can alleviate this issue, such as hybrid word-character-based model (Luong and Manning, 2016), BPE-based model (Sennrich et al., 2016) and word-piece-based model (Wu et al., 2016). These effective work alleviate the token imbalance phenomenon to a certain extent and become the *de-facto* standard in most NMT models. Although sub-word based NMT models have achieved significant improvements, they still face the token-level frequency imbalance phenomenon, as Table 1 shows.

Furthermore, current NMT models generally assign equal training weights to target tokens without considering their frequencies. It is very likely for NMT models to ignore the loss produced by the low-frequency tokens because of their small proportion in the training sets. The parameters related to them can not be adequately trained, which will, in turn, make NMT models tend to prioritize output fluency over translation adequacy, and ignore the generation of low-frequency tokens during decoding, which is illustrated in Table 1. It shows that the vanilla NMT model tends to generate more high-frequency tokens and less low-frequency tokens. However, low-frequency tokens may carry critical semantic information which may affect translation quality once they are neglected.

To address the above issue, we proposed token-level adaptive training objectives based on target token frequencies. We aimed that those meaningful but relatively low-frequency tokens could be assigned with larger loss weights during training so that the model will learn more about them. To explore suitable adaptive objectives for NMT, we first applied existing adaptive objectives from other tasks to NMT and analyzed their performance. We found that though they could bring modest improvement on the translation of low-frequency tokens, they did much damage to the translation of high-frequency tokens, which led to an obvious degradation on the overall performance. This implies that the objective should ensure the training of high-frequency tokens first. Then, based on our observations, we proposed two heuristic criteria for designing the token-level adaptive objectives based on the target token frequencies. Last, we presented two specific forms for different application scenarios according to the criteria. Our method yields consistent improvements in translation quality on ZH-EN, EN-RO, and EN-DE translation tasks, especially on sentences that contain more low-frequency tokens where we can get 1.68, 1.02, and 0.52 BLEU increases compared with baseline, respectively. Further analyses show that our method can also improve the lexical diversity of translation.

Our contributions can be summarized as follows:

- We analyzed the performance of the existing adaptive objectives when they were applied to NMT. Based on our observations, we proposed two heuristic criteria for designing token-level adaptive objectives and present two specific forms to alleviate the problem brought by the token imbalance phenomenon.

- The experimental results validate that our method can improve not only the translation quality, especially on those low-frequency tokens, but also the lexical diversity.

2 Background

In our work, we apply our method in the framework of *Transformer* (Vaswani et al., 2017) which will be briefly introduced here. We denote the input sequence of symbols as $\mathbf{x} = (x_1, \dots, x_J)$, the ground-truth sequence as $\mathbf{y}^* = (y_1^*, \dots, y_K^*)$ and the translation as $\mathbf{y} = (y_1, \dots, y_K)$.

The Encoder & Decoder The encoder is composed of N identical layers. Each layer has two sublayers. The first sublayer is a multi-head attention unit used to compute the self-attention of the input, named *self-attention multi-head sublayer*, and the second one is a fully connected feed-forward network, named *FNN sublayer*. Both of the sublayers are followed by a residual connection operation and a layer normalization operation. The input sequence $\mathbf{x}$ will be first converted to a sequence of vectors $\mathbf{E}_x = [E_x[x_1]; \dots; E_x[x_J]]$, where $E_x[x_j]$ is the sum of the word embedding and the position embedding of the source word x_j. Then, this input sequence of vectors will be fed into the encoder and the output of the N-th layer will be taken as source hidden states. The decoder is also composed of N identical layers. In addition to the same kind of two sublayers in each encoder layer, the third *cross-attention sublayer* is inserted between them, which

	Valid	High	Low
Baseline	45.46	49.27	41.35
Linear	45.33*(-0.13)*	48.59*(-0.68)*	41.64*(+0.29)*
Focal	44.91*(-0.55)*	48.17*(-1.10)*	41.36*(+0.01)*
Focal + 1	45.71*(+0.25)*	49.36*(+0.09)*	41.93*(+0.58)*

Table 2: BLEU on the validation set of the Chinese-English translation task. 'Low' is the subset of the validation set which contains more low-frequency tokens while 'High' contains more high-frequency tokens.

performs multi-head attention over the output of the encoder. The final output of the N-th layer gives the target hidden states $\mathbf{S} = [\mathbf{s}_1; \ldots; \mathbf{s}_I]$, where $\mathbf{s}_i$ is the hidden states of y_k.

The Objective The model is optimized by minimizing a cross-entropy loss with the ground-truth:

$$\mathcal{L} = -\frac{1}{K}\sum_{k=1}^{K} \log p(y_k^* | \mathbf{y}_{<k}, \mathbf{x}), \qquad (1)$$

where K is the length of the target sentence.

3 Method

Our work aims to explore suitable adaptive objectives that can not only improve the learning of low-frequency tokens but also avoid harming the translation quality of high-frequency tokens. We first investigated two existing adaptive objectives, which were proposed for solving the token imbalance problems in other tasks, and analyzed their performance. Then, based on our observations, we introduced two heuristic criteria for designing the adaptive objective. Based on the proposed criteria, we put forward two simple but effective functional forms from different perspectives, which can be adapted to various application scenarios in NMT.

3.1 Existing Adaptive Objectives Investigation

The form of adaptive objective is as follows:

$$\mathcal{L} = -\frac{1}{I}\sum_{i=1}^{I} w(y_i) \log p(y_i | \mathbf{y}_{<i}, \mathbf{x}), \qquad (2)$$

where $w(y_i)$ is the weight assigned to the target token y_i, which varies as the token frequency changes. Actually, there are some existing adaptive objectives which have been proven effective for other tasks. It can help us understand what is necessary for a suitable adaptive objective for NMT if we apply these methods to it. The first objective we have investigated is the form in Focal loss (Lin et al., 2017), which was proposed for solving the label imbalance problem in the object detection task:

$$w(y_i) = (1 - p(y_i))^{\gamma}. \qquad (3)$$

Although it doesn't utilize the frequency information directly, it actually reduces the weights of the high-frequency classes more because they are usually easier to classify with higher prediction probabilities. We set γ to 1 as suggested by their experiments. We noticed that this method greatly reduced the weights of high-frequency tokens, and the variance of weights is large. The second is the linear weighting function (Jiang et al., 2019), which was proposed for the dialogue response generation task:

$$w(y_i) = -\frac{\texttt{Count}(y_i)}{\max(\texttt{Count}(y_k))} + 1, y_k \in \mathrm{V}_t, \qquad (4)$$

where $\texttt{Count}(y_k)$ is the frequency of token y_k in the training set and V_t denotes the target vocabulary. Then, the normalized weights $w(y_i)$, which have a mean of 1, are assigned to the target tokens. We noticed that the weights of high-frequency tokens are only slightly less than 1, and the variance of weights is small. We tested these two objectives on the Chinese to English translation task and the results on the validation set are given in Table 2[1]. To verify their effects on the high- and low-frequency tokens, we also divided the validation set into two subsets based on the average token frequency of the sentences, the results of which are also given in Table 2. It shows that although these two methods can bring modest improvement in the translation of the low-frequency tokens, it does much harm to high-frequency tokens, which has a negative impact on the overall performance. We noted that both of these two methods reduced the weights of the high-frequency tokens to different degrees, and we argued that when the high-frequency tokens account for a large proportion in NMT corpus, this hinders the normal training of them. To validate our argument, we simply add 1 to the weighting term of focal loss:

$$w(y_i) = (1 - p(y_i))^{\gamma} + 1. \qquad (5)$$

The results are also given in Table 2 (Row 5), which indicates that this method actually avoids the damage to the high-frequency tokens. The overall results indicate that it is not robust enough to improve

[1] The details about the data will be given in the experiment section

the learning of low-frequency tokens by reducing the weight of high-frequency tokens during the training of NMT. Although our goal is to improve the training of low-frequency tokens, we should first ensure the training of high-frequency tokens, and then increase the weights of low-frequency tokens appropriately. Based on the above findings, we proposed the following criteria.

3.2 Heuristic Criteria for Token Weighting

We proposed two heuristic criteria for designing the token-level training weights:

Minimum Weight Ensurence. The training weight of any token in the target vocabulary should be equal to or bigger than 1, which can be described as:

$$\forall y_k \in \mathrm{V}_t, w(y_k) \geq 1 \tag{6}$$

Although we can force the model to pay more attention to low-frequency tokens by shrinking the weights of high-frequency tokens, the previous analyses have proved that the training performance is more sensitive to the change of high-frequency tokens' weights due to their large proportion in the training set. A relatively small decrease in the weights of high-frequency tokens will prevent the generation probabilities of ground-truth tokens from ascending continually, which may result in an obvious degradation of the overall performance. Therefore, we ensure that all the token weights are equal to or bigger than 1 considering the training stability as well as designing convenience.

Weights Expectation Range Control. On the condition that the first criterion is satisfied, those high-frequency tokens could have already been well learned without any extra attention. Now, those low-frequency tokens could be assigned with higher weights. Meanwhile, we also need to ensure that the weights of low-frequency tokens can't be too large, or it will hurt the training of high-frequency tokens certainly. Therefore, the expectation of the training weights on the whole training set should be in $[1, 1+\delta]$:

$$\frac{\sum_{k=1}^{|\mathrm{V}_t|} \texttt{Count}(y_k) w(y_k)}{\sum_{k=1}^{|\mathrm{V}_t|} \texttt{Count}(y_k)} = 1 + \delta, \delta \geq 0, \tag{7}$$

where $|\mathrm{V}_t|$ denotes the size of the target vocabulary, δ is a relatively small number compared with 1. A larger weight expectation means we allocate larger weights to those low-frequency tokens. In contrast, an appropriate weight expectation as defined in this criterion can help improve the overall performance.

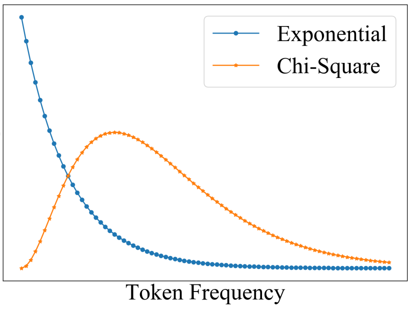

Figure 1: Plots of our two weighting functions. The blue curve is the Exponential form and the orange curve is the Chi-Square form. Both of the hyperparamters are set to 1.

The two criteria proposed here are not the only options for NMT, but the adaptive objective satisfying these two criteria can improve not only the translation performance of low-frequency tokens but also the overall performance based on our experimental observations.

3.3 Two Specific Adaptive Objectives

In this paper, we proposed two simple functional forms for $w(y_k)$ heuristically based on the previous criteria and justified them with some intuitions.

Exponential: Given the target token y_k, we define the exponential weighting function as:

$$w(y_k) = \mathrm{A} \cdot e^{-\mathrm{T} \cdot \texttt{Count}(y_k)} + 1. \tag{8}$$

There are two hyperparameters in it, i.e., A and T, which control the shape and the value range of the function. They can be set up according to the two criteria above. The plot of this weighting function is presented in Figure 1. In this case, we don't consider the factor of noisy tokens so that the weight increases monotonically as the frequency decreases. Therefore, this weighting function is suitable for cleaner training data where the extremely low-frequency tokens only take up a small proportion.

Chi-Square: The exponential form weighting function is not suitable for the training data which contain many noisy tokens, because they would be assigned with relatively large weights and have bigger impacts when their weights are summed together. To alleviate this problem, we proposed another form of the weighting function:

$$w(y_k) = \mathrm{A} \cdot \texttt{Count}^2(y_k) e^{-\mathrm{T} \cdot \texttt{Count}(y_k)} + 1. \tag{9}$$

The form of this function is similar to the form of chi-square distribution, so we named it as chi-square. Plot of this weighting function is presented in Figure 1. We can see from the plot that the weight increases as the frequency decreases at first. Then, after a specific frequency threshold, which is decided by the hyperparameter T, the weight decreases as the frequency decreases. In this case, the most frequent tokens and the extremely rare tokens, which could be noise, all will be assigned with small weights. Meanwhile, those middle-frequency words will have larger weights. Most of them are meaningful and valuable for translation but can't be well learned with an equal-weighted objective function. This form of weighting function is suitable for more noisy training data.

4 Experiments

4.1 Data Preparation

ZH→EN. The training data consists of 1.25M sentence pairs from LDC corpora which has 27.9M Chinese words and 34.5M English words, respectively [2]. The data set MT02 was used as validation and MT03, MT04, MT05, MT06, MT08 were used for the test. We tokenized and lowercased English sentences using the Moses scripts[3], and segmented the Chinese sentences with the Stanford Segmentor[4]. The two sides were further segmented into subword units using Byte-Pair Encoding (BPE) (Sennrich et al., 2016) with 30K merge operations separately.

EN→RO. We used the preprocessed version of the WMT2016 English-Romanian dataset released by Lee et al. (2018) which includes 0.6M sentence pairs. We used news-dev 2016 for validation and news-test 2016 for the test. The two languages shared the same vocabulary generated with 40K merge operations of BPE.

EN→DE. The training data is from WMT2016 which consists of about 4.5M sentences pairs with 118M English words and 111M German words. We chose the news test-2013 for validation and news-test 2014 for the test. 32K merge operations BPE were performed on both sides jointly.

4.2 Systems

We used the open-source toolkit called *Fairseq-py* (Edunov et al., 2017) released by Facebook as our Transformer system.

- **Baseline**. The baseline system was implemented as the base model configuration in Vaswani et al. (2017) strictly. Since our method is further trained based on the pre-trained model at a low learning rate, we also trained another baseline model following the same procedures as our methods have except that all the target tokens share equal weights in the objective, denoted as **Baseline-FT**.
- **Fine Tuning** (Luong and Manning, 2015). This model was first trained with all the training sentence pairs and then further trained with sentences containing more low-frequency tokens. To filter out sentences containing more low-frequency tokens, the method in Platanios et al. (2019) was adopted as our judging metric with a small modification:

$$d_{\text{rarity}}(\mathbf{y}) \triangleq -\frac{1}{I}\sum_{i=1}^{I}\log\frac{\texttt{Count}(y_i)}{\sum_{k=1}^{|\mathrm{V}_t|}\texttt{Count}(y_k)}, \tag{10}$$

where I is the sentence length. We added a factor $\frac{1}{I}$ to eliminate the influence of sentence length. All the target sentences were ranked by this metric in ascending order and the bottom one third of the training sentences were chosen as the in-domain data. This method tries to utilize frequency information at the sentence level, while our work uses it at the token level in contrast.

- **Sampler** (Chu et al., 2017). This method oversampled the sentences containing more low-frequency tokens filtered by Eq. 10 three times and then concatenated them with the rest of the training data. Thus the NMT model will be trained with more low-frequency tokens in every epoch.
- **Entropy Regularization (ER)** (Pereyra et al., 2017). This method was proposed for solving the overconfidence problem, which adds a confidence penalty term to the original objective:

$$\mathcal{L}_{\text{ER}} = L - \alpha\frac{1}{I}\sum_{i=1}^{I}p(y_i|\mathbf{x})\log(p(y_i|\mathbf{x})). \tag{11}$$

It is known that token imbalance is one of the causes of overconfidence problem (Jiang and de Rijke, 2018), so this method may also alleviate the token imbalance problem. We varied α from 0.05 to 0.4 and chose the best one according to the results on the validation sets for different languages. Noting that the label smoothing is applied in the vanilla transformer model which has a similar effect on the output, we removed it from the model when we tested this method.

[2] The corpora include LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06.

[3] http://www.statmt.org/moses/

[4] https://nlp.stanford.edu/

	T	ZH-EN	EN-RO	EN-DE
Baseline	-	45.49	33.60	25.45
Our_Exp	0.25	46.07	-	-
	0.35	**46.28**	-	-
	0.50	46.19	34.10	-
	0.75	46.13	34.11	-
	1.00	46.01	34.24	26.02
	1.25	-	**34.26**	26.01
	1.50	-	34.15	26.06
	1.75	-	34.15	**26.10**
	2.00	-	-	26.03
Our_K2	1.50	46.14	-	-
	1.75	**46.24**	-	-
	2.00	46.00	34.07	-
	2.50	45.98	-	**26.06**
	3.00	-	34.07	25.93
	4.00	-	**34.15**	25.87
	5.00	-	34.10	25.95

Table 3: Performance of our methods on the validation sets for all the three language pairs with different hyperparameters T. Although the best hyperparameter for different languages may be different, it is easy for our method to get a stable improvement.

- **Linear** (Jiang et al., 2019). This method was proposed for solving the token imbalance problem in the the dialogue response generation task:

$$w(y_i) = -\frac{\texttt{Count}(y_i)}{\max(\texttt{Count}(y_k))} + 1, y_k \in \mathrm{V}_t. \quad (12)$$

Then, the normalized weights, which had a mean of 1, were applied to the training objective.

- **Our_Exp**. This system was first trained with the normal objective (Equation 1), where all the target tokens have the same training weights. Then the model was further trained with the adaptive objective at a low learning rate. The weights were produced by the Exponential form (Equation 8). For computing stability, we used $\frac{\texttt{Count}(y_k)}{\mathrm{C}_{\text{median}}}$ instead of $\texttt{Count}(y_k)$ in the weighting function, where $\mathrm{C}_{\text{median}}$ is the median of the token frequency.
- **Our_K2**. This system was trained following the same procedure as system Our_Exp except that the training weights were produced by the Chi-Square form (Equation 9).

The translation quality was evaluated by 4-gram BLEU (Papineni et al., 2002) with the *multi-bleu.pl* script. Besides, we used beam search with a beam size of 4 and a length penalty of 0.6 during the decoding process.

4.3 Hyperparameters

There are two hyperparameters in our weighting functions, A and T. In our experiments, we fixed A to narrow search space and the overall weight range is $[1, e]$. We tuned another hyperparameter T on the validation data sets under the criteria proposed in section 3.2. The results are shown in Table 3. According to the results, the best hyperparameters differed across different language pairs. It is affected by the proportion of low-frequency words and high-frequency words. Generally speaking, when the proportion of low-frequency words gets smaller, the hyperparameter T should be set smaller too. But it also shows that it is easy for our methods to get a stable improvement over the baseline system following the criteria above. Finally, we used the best hyperparameters as found on the validation data sets for the final evaluation of the test data sets. For example, $\mathrm{T} = 0.35$ in the exponential form for ZH→EN and $\mathrm{T} = 4.00$ in the chi-square form for EN→RO.

4.4 Main Results

The results are shown in Table 4. It shows that the contrast methods can not bring stable improvements over the baseline system. They bring excessive damages to the translation of high-frequency tokens which can be proved by the analyzing experiments in the next section. As a contrast, our methods can bring stable improvements over Baseline-FT almost without any additional computing or storage expense. On the EN→RO and EN→DE translation tasks, Our_Exp is more effective than Our_K2 while on the ZH→EN translation task the result is reversed. The reason is that the NIST training data set contains more noisy tokens, which can be ignored by the Our_K2 method. More analyses based on the token frequency are shown in the next section.

5 Analysis

5.1 Effects on Translation Quality with Considering Token Frequencies

To further illustrate the effects of our method, we evaluated the performance based on the token frequency. For the ZH→EN translation task, we concatenated the MT03-08 test sets together as a big test set. For the EN→RO and EN→DE translation tasks, we just used their test sets. Each sentence was scored according to Eq. 10 and sorted in ascending order. Then the test set was divided into

	ZH→EN							EN→RO		EN→DE	
	MT03	MT04	MT05	MT06	MT08	AVE	Δ	WMT16	Δ	WMT16	Δ
Baseline	44.63	45.79	44.03	43.78	35.63	42.77		32.85		27.15	
Baseline-FT	44.69	46.24	44.01	44.33	35.83	43.02		33.15		27.21	
Fine Tuning	45.06	46.30	45.30	43.61	34.68	42.99	*-0.03*	33.28	*+0.13*	26.56	*-0.65*
Sampler	44.85	46.02	44.57	44.04	35.02	42.90	*-0.12*	32.75	*-0.40*	-	-
ER	44.31	46.38	45.13	44.29	35.71	43.16	*+0.14*	33.21	*+0.06*	27.19	*-0.02*
Linear	44.26	46.02	43.99	44.08	34.71	42.62	*-0.60*	33.35	*+0.20*	27.37	*+0.16*
Our_Exp	45.67**	47.02**	45.43**	44.51	36.11	43.75	*+0.73*	**33.77****	***+0.62***	**27.60****	***+0.39***
Our_K2	**45.87****	**47.07****	**45.62****	**44.72**	**36.20**	**43.90**	***+0.88***	33.54*	*+0.49*	27.51*	*+0.30*

Table 4: BLEU scores on three translation tasks. The column of Δ shows the improvement compared to Baseline-FT. ** and * mean the improvements over Baseline-FT is statistically significant (Collins et al., 2005) ($\rho < 0.01$ and $\rho < 0.05$, respectively). The results show that our methods can achieve significant improvements on translation quality.

	ZH→EN			EN→RO		
	High	Middle	Low	High	Middle	Low
Baseline-FT	50.88	43.06	34.90	35.68	33.61	29.86
Fine Tuning	49.85(*-1.03*)	42.68(*-0.38*)	35.85(*+0.95*)	35.51(*-0.17*)	33.45(*-0.16*)	30.56(*+0.70*)
Sampler	49.77 (*-1.11*)	42.63(*-0.43*)	35.77(*+0.87*)	35.22(*-0.46*)	33.07(*-0.54*)	30.10(*+0.42*)
ER	50.59 (*-0.29*)	42.82(*-0.25*)	35.48(*+0.58*)	35.66(*-0.03*)	33.25(*-0.36*)	30.26(*+0.41*)
Linear	50.21 (*-0.67*)	43.06(*-0.68*)	35.19(*+0.29*)	35.57(*-0.11*)	33.65(*+0.04*)	30.35(*+0.49*)
Our_Exp	50.88(*+0.00*)	43.30(*+0.24*)	36.45**(*+1.55*)	**36.08(*+0.40*)**	**34.26*(*+0.65*)**	**30.88**(*+1.02*)**
Our_K2	**51.07(*+0.19*)**	**43.31(*+0.25*)**	**36.58**(*+1.68*)**	35.94(*+0.26*)	33.97(*+0.36*)	30.65**(*+0.79*)

Table 5: BLEU scores on different test subsets which are grouped by their rarities according to Eq. 10. Sentences in the 'Low' contain more low-frequency tokens while the 'High' is reverse. The results show that our methods can improve the translation of low-frequency tokens significantly without hurting the translation of high-frequency tokens.

	High	Middle	Low
Baseline-FT	28.88	26.97	25.55
Fine Tuning	26.40(*-2.48*)	26.69(*-0.28*)	25.84(*+0.29*)
ER	28.72(*-0.16*)	26.86(*-0.11*)	25.74(*+0.19*)
Linear	28.88(*+0.00*)	27.07(*+0.10*)	25.70(*+0.15*)
Our_Exp	**28.91(*+0.03*)**	**27.33*(*+0.36*)**	**26.07**(*+0.52*)**
Our_K2	28.90(*+0.02*)	27.28*(*+0.31*)	25.99*(*+0.44*)

Table 6: EN→DE BLEU scores on different test subsets. The conclusion is identical to that in Table 5.

three subsets with equal size, denoted as High, Middle, and Low, respectively. Sentences in the subset Low contain more low-frequency tokens while the High is reverse.

The results are given in Table 5 and Table 6. The contrast methods outperform the Baseline-FT on the Low subset but are worse than it in the High and Middle subsets, which indicates that the gains on the translation of low-frequency tokens come at the expense of the translation of high-frequency tokens. As a contrast, both of our methods can not only bring a significant improvement on the Low subset but also get a modest improvement on the High and Middle subsets. It can be concluded that our methods can ameliorate the translation of low-frequency tokens without hurting the translation of high-frequency tokens.

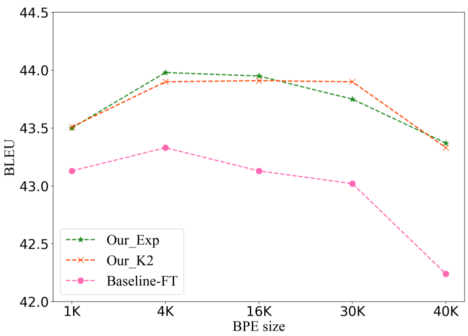

Figure 2: BLEU with different BPE sizes on ZH→EN translation task. It shows that our method can always bring a stable improvement compared with the baseline.

5.2 Effects on Translation Quality with Different BPE Sizes

It is known that the BPE sizes have a large impact on the data distribution. Intuitively, a smaller size of BPE will bring a more balanced data distribu-

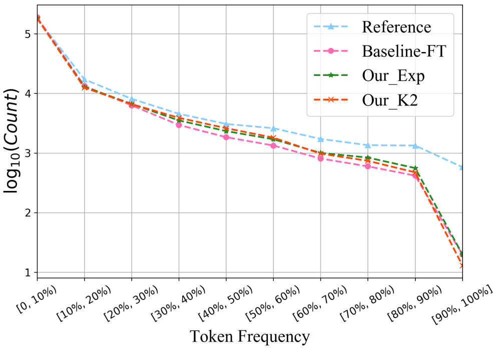

Figure 3: The count of tokens with different frequencies in references, translations of the baseline systems and our methods on the ZH→EN translation task. The tokens are ranked by their frequencies in the training sets. The x-axis represents the frequency interval ([20%, 30%) means the frequency of tokes is between top 20% and 30%), the y-axis is the count of the tokens applied with a common logarithm operation in each interval.

	TTR($\times 10^{-2}$)	HD-D	MTLD
Baseline-FT	5.32	0.829	59.1
Our_Exp	5.87	0.836	62.2
Our_K2	5.95	0.835	61.9
Reference	6.79	0.852	69.2

Table 7: The lexical diversity of translations. A larger value represents higher diversity. The results show that our method can improve the lexical diversity.

tion, but it will also increase the average sentence length and neglect some token co-occurrences. To verify the effectiveness of our method with different BPE sizes, we varied the BPE sizes from 1K to 40K on the ZH→EN translation task. The results are shown in Figure 2. It shows that as the number of BPE size increases, the BLEU of baseline rises first and then declines. Compared with the baseline systems, our method can always bring improvements, and the larger the BPE size, i.e., the more imbalanced the data distribution, the larger the improvement brought by our method. In practice, the BPE size either comes from the experience or is chosen from several trial-and-errors. No matter what the situation is, our method can always bring a stable improvement.

5.3 Effects on Token Distribution and Lexical Diversity

Compared with the reference, the outputs of the vanilla NMT model contain more high-frequency tokens and have lower lexical diversity (Vanmassenhove et al., 2019b). To verify whether our methods can alleviate these problems, we did the following experiments based on the ZH→EN translation task. The tokens in the target vocabulary were first arranged in descending order according to their token frequencies. Then they were divided into ten intervals equally. Finally, we counted the number of tokens in each token frequency interval of the reference and the translation of different systems. The results are shown in Figure 3 and we did a common logarithm for display convenience. It shows that there is an obvious gap between the Baseline-FT and reference, and the curve of Baseline-FT is lower than the curve of reference in every frequency interval except for the top 10%. As a contrast, our methods can reduce this gap, and the tokens distribution is closer to the real distribution. Besides, we also measure the lexical diversity of the translations with several criteria, namely, type-token ratio (TTR) (Templin, 1957), the approximation of hypergeometric distribution (HD-D) and the measure of textual lexical diversity (MTLD) (Mccarthy and Jarvis, 2010). The results are given in Table 7. It shows that our method can also improve the lexical diversity of the translation.

Source	búduàn guānbì nàxiē wūrǎn huánjìng de **méikuàng** .
Reference	those **coalmines** pollute the environment should be continuously shut down .
Baselie-FT	continually close down those **mines** that pollute the environment .
Our_Exp	those **coalmines** that pollute the environment should be continuously closed.
Our_K2	those **coalmines** that pollute the environment should be continuously closed.
Source	yǐhòu kěyǐ gěi wǒ dāndú **pèi** jiān bàngōngshì .
Reference	an exclusive office could be **assigned** me later on .
Baselie-FT	later i could **match** my office alone .
Our_Exp	i could be **assigned** an office alone later .
Our_K2	later i could be **assigned** an office alone .

Table 8: Translation examples of the Basline-FT and our methods. The results show that our methods can generate low-frequency but more accurate tokens.

5.4 Case Study

Table 8 shows two translation examples in the ZH→EN translation direction. In the first sentence, the Baseline-FT system failed to generate the low-frequency noun '*coalmine*' (frequency: 43), but generated a relatively high-frequency word '*mine*' (frequency: 1155). We can see that this low-frequency token carries the central information of this sentence, and the mistranslation of it prevents people from understanding this sentence correctly. In the second sentence, our methods generated the low-frequency verb '*assigned*' (frequency: 841) correctly, while the Baseline-FT generated a more frequent token '*match*' (frequency: 1933), which reduced the translation accuracy and fluency. These examples can be part of the evidence to show the effectiveness of our methods.

6 Related Work

Rare Word Translation. Rare word translation is one of the key challenges for NMT. For word-level NMT models, NMT has its limitation in handling a larger vocabulary because of the training complexity and computing expense. Some work tries to solve this problem by maintaining phrase tables or back-off vocabulary (Luong et al., 2015; Jean et al., 2015; Li et al., 2016). The subword-based NMT (Sennrich et al., 2016; Luong and Manning, 2016; Wu et al., 2016) reduces the size of vocabulary greatly and become the mainstream technology gradually. Gowda and May (2020) gave a detailed analysis about the effects of the BPE size on the data distribution and translation quality. Some recent work tried to further improve the translation of the rare words with the help of the memory network or the pointer network (Zhao et al., 2018; Pham et al., 2018). In contrast, our methods can improve the translation performance without extra cost and can be combined with other techniques.

Class Imbalance. Class imbalance means the total number of some classes of data is far less than the total number of other classes. This problem can be observed in various tasks (Wei et al., 2013; Johnson and Khoshgoftaar, 2019). In NMT, the class imbalance problem might be the underlying cause of, among others, the gender-biased output problem (Vanmassenhove et al., 2019a), the inability of MT system to handle morphologically richer language correctly (Passban et al., 2018), or the exposure bias problem (Ranzato et al., 2016; Shao et al., 2018; Zhang et al., 2019). The methods of trying to solve this can be divided into two types. The data-based methods (Baloch and Rafi, 2015; Ofek et al., 2017) make use of over- and under-sampling to reduce the imbalance. The algorithm-based methods (Zhou and Liu, 2005; Lin et al., 2017) give extra reward to different classes. Our method is algorithm-based which brings no extra cost.

Word Frequency-based Methods. Some work also makes use of word frequency information to help learning, such as in the word segmentation (Sun et al., 2014) and term extraction (Frantzi et al., 1998; Vu et al., 2008). In NMT, word frequency information is used for curriculum learning (Kocmi and Bojar, 2017; Zhang et al., 2018; Platanios et al., 2019) and domain adaptation data selection (Wang et al., 2017; Zhang and Xiong, 2018; Gu et al., 2019). Wang et al. (2020) analyzed the miscalibration problem on the low-frequency tokens. Jiang et al. (2019) proposed a linear weighting function to solve the word imbalance problem in the dialogue response generation task. Compared with it, our method is more suitable for NMT.

7 Conclusion

In this work, we focus on the token imbalance problem of NMT. We show that the output of vanilla NMT contains more high-frequency tokens and has lower lexical diversity. To alleviate this problem, we investigated existing adaptive objectives for other tasks and then proposed two heuristic criteria based on the observations. Next, we gave two simple but effective forms based on the criteria, which can assign appropriate training weights to target tokens. The final results show that our methods can achieve significant improvement in performance, especially on sentences that contain more low-frequency tokens. Further analyses show that our method can also improve the lexical diversity.

Acknowledgements

We thank all the anonymous reviewers for their insightful and valuable comments. This work was supported by National Key R&D Program of China NO. 2017YFE0192900).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR*.

Maher Baloch and Muhammad Rafi. 2015. An investigation on topic maps based document classification with unbalance classes. *Journal of Independent Studies and Research*, 13(1):50.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1724–1734.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *arXiv preprint arXiv:1701.03214*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 531–540. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, and Sam Gross. 2017. https://github.com/pytorch/ fairseq.

Katerina T Frantzi, Sophia Ananiadou, and Junichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *International conference on theory and practice of digital libraries*, pages 585–604. Springer.

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

Thamme Gowda and Jonathan May. 2020. Neural machine translation with imbalanced classes. *CoRR*, abs/2004.02334.

Shuhao Gu, Yang Feng, and Qun Liu. 2019. Improving domain adaptation translation with domain invariant and specific information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3081–3091.

Çaglar Gülçehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Sébastien Jean, KyungHyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1–10.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2879–2885.

Shaojie Jiang and Maarten de Rijke. 2018. Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots. In *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 81–86.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Tom Kocmi and Ondrej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 379–386.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2852–2858. AAAI Press.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural*

Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, pages 11–19.

Philip M Mccarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Nir Ofek, Lior Rokach, Roni Stern, and Asaf Shabtai. 2017. Fast-cbus: A fast clustering-based undersampling method for addressing the class imbalance problem. *Neurocomputing*, 243:88–102.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Peyman Passban, Andy Way, and Qun Liu. 2018. Tailoring neural architectures for translating from morphologically rich languages. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3134–3145.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.

Ngoc-Quan Pham, Jan Niehues, and Alexander H. Waibel. 2018. Towards one-shot learning for rare-word translation with external experts. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, NMT@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 100–109.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom M. Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1162–1172.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*.

Chenze Shao, Xilin Chen, and Yang Feng. 2018. Greedy search with probabilistic n-gram matching for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4778–4784.

Xu Sun, Wenjie Li, Houfeng Wang, and Qin Lu. 2014. Feature-frequency–adaptive on-line training for fast and accurate natural language processing. *Computational Linguistics*, 40(3):563–586.

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

Mildred C. Templin. 1957. Certain language skills in children: Their development and interrelationships.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019a. Getting gender right in neural machine translation. *CoRR*, abs/1909.05088.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019b. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*, pages 222–232.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 5998–6008.

Thuy Vu, Aiti Aw, and Min Zhang. 2008. Term extraction through unithood and termhood unification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1482–1488.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3070–3079.

Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, and Jiahang Chen. 2013. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16(4):449–475.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Shiqi Zhang and Deyi Xiong. 2018. Sentence weighting for neural machine translation domain adaptation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3181–3190.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4334–4343.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J. Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *CoRR*, abs/1811.00739.

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

Zhi-Hua Zhou and Xu-Ying Liu. 2005. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77.

George Kingsley Zipf. 1949. Human behavior and the principle of least effort.