Dual-teacher Knowledge Distillation for Low-frequency Word Translation

Yifan Guo Hongying Zan Hongfei Xu*

Zhengzhou University, Henan, China
yfguonlp@foxmail.com, iehyzan@zzu.edu.cn, hfxunlp@foxmail.com

Abstract

Neural Machine Translation (NMT) models are trained on parallel corpora with unbalanced word frequency distribution. As a result, NMT models are likely to prefer high-frequency words than low-frequency ones despite low-frequency word may carry the crucial semantic information, which may hamper the translation quality once they are neglected. The objective of this study is to enhance the translation of meaningful but low-frequency words. Our general idea is to optimize the translation of low-frequency words through knowledge distillation. Specifically, we employ a low-frequency teacher model that excels in translating low-frequency words to guide the learning of the student model. To remain the translation quality of high-frequency words, we further introduce a dual-teacher distillation framework, leveraging both the low-frequency and high-frequency teacher models to guide the student model's training. Our single-teacher distillation method already achieves a $+0.64$ BLEU improvements over the state-of-the-art method on the WMT 16 English-to-German translation task on the low-frequency test set. While our dual-teacher framework leads to $+0.87$, $+1.24$, $+0.47$, $+0.87$ and $+0.86$ BLEU improvements on the IWSLT 14 German-to-English, WMT 16 English-to-German, WMT 15 English-to-Czech, WMT 14 English-to-French and WMT 18 Chinese-to-English tasks respectively compared to the baseline, while maintaining the translation performance of high-frequency words.

1 Introduction

Neural machine translation models typically require large amounts of parallel corpora (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Bahdanau et al., 2014; Sutskever et al., 2014; Gehring et al., 2017; Vaswani et al., 2017). While such data normally have an unbalanced word distribution, the translation models trained on the data usually tend to favor high-frequency words while ignoring low-frequency words. Gu et al. (2020) point out the severe imbalance issue between high-frequency and low-frequency words, and that translation models rarely have the opportunity to learn the true labels of low-frequency words during training. As a result, NMT models rarely have the opportunity to learn and generate those ground truth low-frequency tokens, even though these low-frequency words often carry important semantic information, typically representing specific concepts or emotions found in certain domains, literary work, or dialects.

To improve the translation of rare words, Luong et al. (2015); Jean et al. (2015); Li et al. (2016); Pham et al. (2018) maintain a phrase table or low-frequency word table, and Gulçehre et al. (2016); Zhao et al. (2018) introduce additional components to the model. However, these approaches brought additional inference complexity and computational costs. The imbalance word distribution issue can be alleviated by segmenting low-frequency sub-words into high-frequency ones while applying Byte Pair Encoding (BPE) (Sennrich et al., 2016; Wu et al., 2016), but the problem remains. Gu et al. (2020) explore target token-level adaptive objectives based on token frequencies to assign larger weights to meaningful but relatively low-frequency words.

Li et al. (2021) have shown that knowledge distillation is effective for long-tailed visual recognition. In this paper, we utilize knowledge distillation to optimize the translation of low-frequency words. We obtain a low-frequency teacher model by fine-tuning on the low-frequency part of the training set. Then we use knowledge distillation to guide the learning of the student model for low-frequency word translation while retaining its performance on high-frequency words. Furthermore, we propose using dual teacher models to guide the student model in learning both high-frequency and low-frequency words, to further ensure the translation

*Corresponding author: Hongfei Xu.

Findings of the Association for Computational Linguistics: EACL 2024, pages 5543–5552

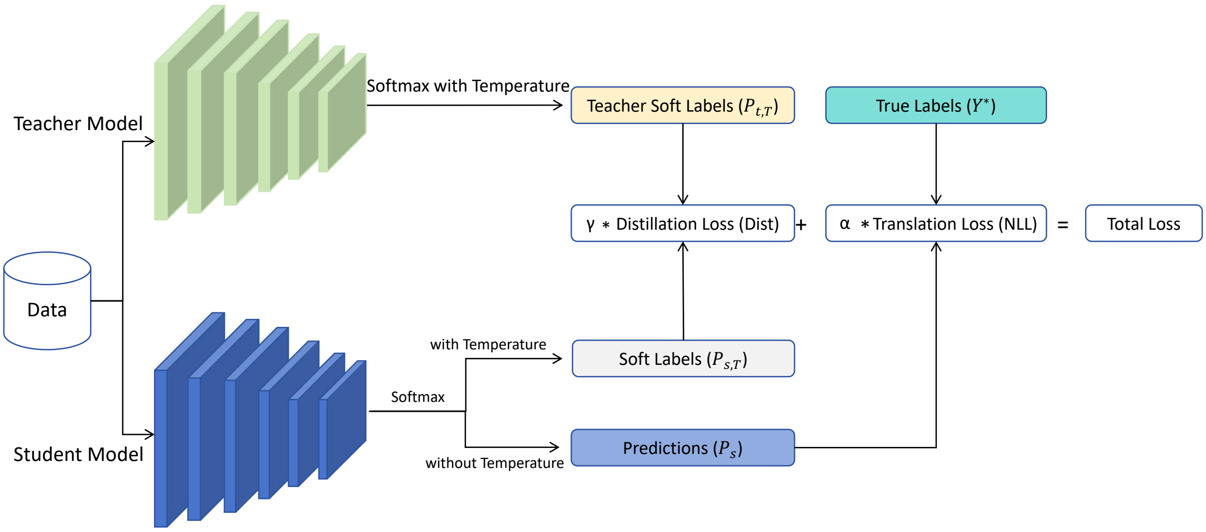

Figure 1: Knowledge distillation with low-frequency teacher model.

performance of high-frequency words of the student model. Our main contributions are as follows:

- We propose to improve the performance of low-frequency word translation through knowledge distillation, transferring the translation knowledge of low-frequency words effectively from the low-frequency teacher model to the student model.

- To further ensure the performance of high-frequency word translation on very large datasets, we introduce a dual-teacher knowledge distillation framework. It utilizes two teacher models to simultaneously guide the learning of both high-frequency and low-frequency words.

- Our single-teacher distillation method already achieves +0.64 BLEU improvements over the state-of-the-art method on the WMT 16 English→German translation task on the low-frequency test set without hamper the performance on the high-frequency test set. While our dual-teacher framework leads to +0.87, +1.24, +0.47, +0.87 and +0.86 BLEU improvements on the IWSLT 14 German→English, WMT 16 English→German, WMT 15 English→Czech, WMT 14 English→French and WMT 18 Chinese→English tasks respectively compared to the baseline, while maintaining the performance on the high-frequency test set even on very large datasets.

2 Our Method

2.1 Low-frequency Word Translation based on Knowledge Distillation

We fine-tune the NMT model on the low-frequency part of the training set to obtain the low-frequency teacher model, and use the prediction probability of the teacher model to supervise the training of the student model together with the original translation loss, as shown in Figure 1.

For the input sentence $X = (x_1, x_2, ..., x_n)$ and the corresponding target translation $Y^* = (y_1, y_2, ..., y_m)$ in a training instance. The Transformer encoder takes the the sum of the corresponding word vectors of X and position encodings as input, and transforms it into a sequence of contextual representations.

The output of the encoder is fed into the decoder for the computation of cross-attention layers. The output of the last decoder layer $H_{dec} = [[H_{dec,1}], [H_{dec,2}], ..., [H_{dec,m}]]$ is to predict the probability of each token with the softmax function.

However, the probabilities of many tokens are close to zero after softmax, especially with the large vocabulary size of the machine translation task. It can be difficult for the student model to learn from the probability distribution which is almost full-filled very small probabilities. To address this issue, we employ a temperature hyper-parameter T to smooth the probability distribution following Hinton et al. (2015), as shown in Equation 1.

$$Out_T = Softmax(\frac{\mathbf{Y}_s}{T}) \quad (1)$$

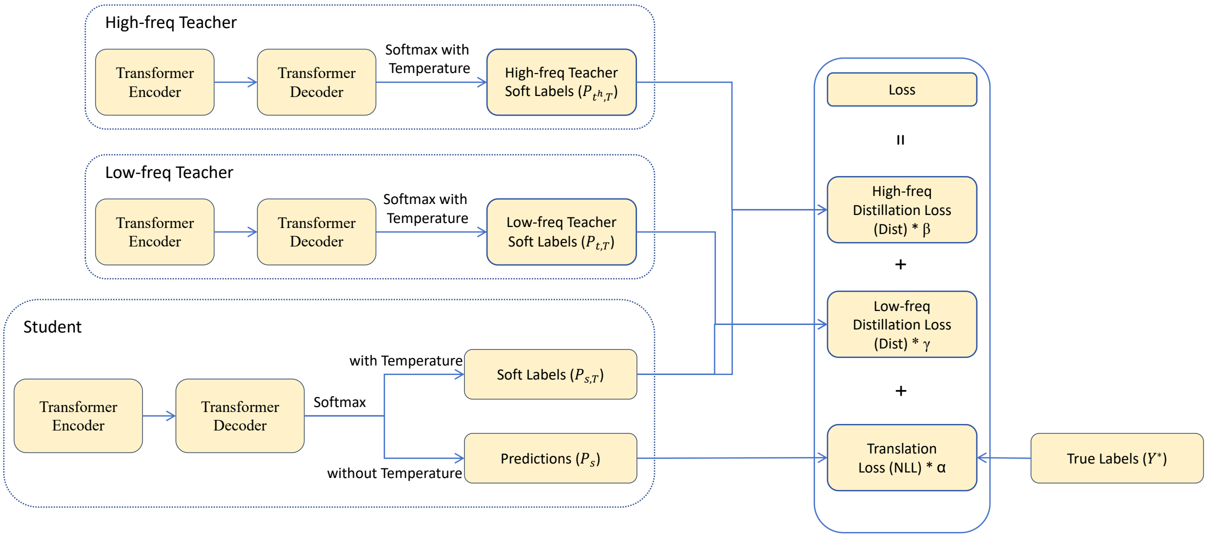

Figure 2: Dual-teacher knowledge distillation teacher model guides student model.

The training loss is the weighted combination of the original translation loss and the knowledge distillation loss with α and γ as corresponding weights. The knowledge distillation loss is computed by minimize the distance function $Dist$ between the probability distribution of the teacher model with temperature $P_{t,T}$ and that of the student model $P_{s,T}$. The machine translation loss is computed by minimize the negative log likelihood loss NLL given the probability of the student model without temperature P_s and the reference translation Y^*, as shown in Equation 2.

$$Loss = \alpha * NLL(P_s, Y^*) + \gamma * Dist(P_{s,T}, P_{t,T}) \quad (2)$$

2.2 Dual-teacher Knowledge Distillation

With the growing size of the training set and the limited capacity of the student model, distilling only with the low-frequency teacher may take up the capacity for high-frequency word translation and deteriorate the performance of high-frequency words on very large datasets. We employ a high-frequency teacher in addition to the low-frequency teacher to preserve the translation quality of high-frequency words while improving that of the low-frequency words. The dual-teacher distillation framework is shown in Figure 2.

The knowledge distillation loss for high frequency words optimizes the probability distribution distance between prediction probability distribution $P_{t^h,T}$ of the high-frequency teacher model t^h and that of the student model. The training loss is the weighted combination of the machine translation loss and low-frequency and high-frequency distillation losses with α, γ and β as corresponding weights, as shown in Equation 3.

$$\begin{aligned} Loss = \alpha * NLL(P_s, Y^*) + \\ \gamma * Dist(P_{s,t}, P_{t,T}) + \\ \beta * Dist(P_{s,t}, P_{t^h,T}) \end{aligned} \quad (3)$$

2.3 Distillation Loss

Huang et al. (2022) show that when a more powerful teacher model exhibits significant differences from the student model in knowledge distillation, the performance of the student model may decline, and can even be worse than training from scratch without knowledge distillation. To address this, Huang et al. (2022) propose a method that focuses only on the preferences of the teacher model with pearson correlation, which refers to the relative ranking of predicted results. Instead of asking the student model to exactly mimic absolute values with the Kullback-Leibler (KL) divergence loss, pearson correlation focuses on the relative relationships between different categories predicted by the teacher model.

The Pearson's distance metric d_p is shown in Equation 4.

$$d_p(\mathbf{u}, \mathbf{v}) = 1 - \rho(\mathbf{u}, \mathbf{v}) \quad (4)$$

where $\rho(\mathbf{u}, \mathbf{v})$ is the Pearson correlation coefficient between two random variables u and v.

The computation of the Pearson correlation coefficient is based on the the covariance $Cov(\mathbf{u}, \mathbf{v})$ of $\mathbf{u}$ and $\mathbf{v}$ and their standard derivations, as shown in Equation 5.

$$\rho(\mathbf{u},\mathbf{v}) = \frac{Cov(\mathbf{u},\mathbf{v})}{Std(\mathbf{u})Std(\mathbf{v})} = \frac{\sum_{i=1}^{C}(u_i-\bar{u})(v_i-\bar{v})}{\sqrt{\sum_{i=1}^{C}(u_i-\bar{u})^2\sum_{i=1}^{C}(v_i-\bar{v})^2}} \quad (5)$$

where $\bar{u}$ and $Std(\mathbf{u})$ denote the mean and standard derivation of $\mathbf{u}$ respectively.

By optimizing the pearson correlation instead of the KL divergence, the learning difficulty of the student model regarding the teacher model is effectively reduced, resulting in more stable distillation results.

3 Experiment

3.1 Settings

We conducted our experiments on the following tasks to test the effectiveness of our approach:

- **IWSLT 2014 De→En** To evaluate the performance of the model on low-resource datasets, we selected the German-English dataset from IWSLT 2014. The training data consists of about 174K sentences pairs.
- **WMT 2016 En→De** The training set contains approximately 4.5M sentence pairs. The validation set and test set are newstest 2013 and newstest 2014 respectively.
- **WMT 2015 En→Cs** The training data consists of about 10M sentences pairs. We chose newstest 2013 and newstest 2015 as the validation and test sets respectively.
- **WMT 2018 Zh→En** To evaluate the applicability of the model across different regional languages, we utilized a preprocessed Chinese-English dataset from WMT18. The training set consists of about 19M sentences pairs.
- **WMT 2014 En→Fr** This task is chosen to test performance on large-scale datasets. The training data is from WMT 2014 which consists of about 36M sentence pairs. We chose newstest 2013 and newstest 2014 as the validation and test sets respectively.

We tokenized and truecased sentences using the Moses scripts for all languages except Chinese, and applied shared Byte-Pair Encoding (BPE) with 32K merge operations to address the unknown word issue for the WMT 2016 EN→DE, WMT 2015 EN→CS and WMT 2014 EN→FR tasks, shared BPE with $16k$ merge operations for the low-resource IWLST 2014 DE→EN task, independent BPE with $32k$ merge operations for the WMT 2018 Zh→EN task.

Following Gu et al. (2020), we score data instances of the training set and test set based on word frequencies using Equation 6.

$$Freq_{sentence} = -\frac{1}{L}\sum_{i=0}^{L} log\frac{Count(y_i)}{\sum_{k=1}^{|V_t|}Count(y_k)} \quad (6)$$

where L represents the sentence length, and $\frac{1}{L}$ is to eliminate the influence of sentence length. $Count(y_i)$ represents the frequency of word y_i in the sentence, while $Count(y_k)$ represents the frequency of word y_k in the training set.

A higher score for a sentence indicates that the sentence contains more low-frequency words. After sorting the training set and test set according to the scores, we divided them into three parts of equivalent number of sentence pairs, denoted as $\{Train_{high}, Train_{middle}, Train_{low}\}$ and $\{Test_{high}, Test_{middle}, Test_{low}\}$.

We followed the Transformer Base setting of Vaswani et al. (2017) for all tasks except for the low-resource IWSLT 2014 De→En. We adopted the Transformer with 6 encoder and decoder layers, 512 as the embedding dimension and 4 times of embedding dimension as the number of hidden units of the feed-forward layer, a dropout probability of 0.1. The number of warm-up steps was set to $8k$. We used a batch size of around $25k$ target tokens achieved by gradient accumulation, and trained the models for $100k$ steps. For the low-resource IWSLT 2014 De→En, we followed the experiment settings of Araabi and Monz (2020).

As the student model of knowledge distillation is initialized with the converged Transformer Base model, we also fine-tune the converged base model for another $100k$ training steps to obtain the BaseFT model as our baseline for fair comparison. The learning rate for both knowledge distillation and BaseFT's fine-tuning is 10^{-5}.

To obtain the teacher model that has better performance on low-frequency words, we fine-tuned the converged base model on the low-frequency part of the training set $Train_{low}$. The performance of the low-frequency teacher gets improved on the

	$Test_{low}$	$Test_{middle}$	$Test_{high}$
Transformer (Vaswani et al., 2017)	25.17	26.72	28.56
BaseFT	25.55	26.97	28.88
Token Weighting			
SPL (Wan et al., 2020)	24.46	27.60	31.12
BMI (Xu et al., 2021)	24.08	27.33	30.99
CMBI (Zhang et al., 2022)	23.96	27.60	31.20
SE (Peng et al., 2023)	24.90	27.53	31.07
General KD			
SKD (Wang et al., 2021)	24.84	27.86	31.51
TIE-KD (Zhang et al., 2023a)	25.17	28.10	31.50
Low-Frequency Word Translation			
ER (Pereyra et al., 2017)	25.74	26.86	28.72
Linear (Jiang et al., 2019)	25.70	27.07	28.88
Exponential (Gu et al., 2020)	26.07	27.33	28.91
Chi-Square (Gu et al., 2020)	25.99	27.28	28.90
Ours			
Single teacher	26.71	27.40	28.97
Dual teacher	**26.79**	27.44	28.99

Table 1: Main results on the WMT 16 English→German task.

low-frequency test set but decreased on the high-frequency test set compared to the converged base model. We used BaseFT as the high-frequency teacher model, as the fine-tuning further boosts the performance of the converged base model on high-frequency words.

In addition to the vanilla Transformer and the BaseFT model which fine-tunes the pre-trained Transformer for another $100k$ training steps, we compare our method with a series of baselines related to token weighting (Self-Paced Learning (SPL) (Wan et al., 2020), adaptive training based on Bilingual Mutual Information (BMI) (Xu et al., 2021) and Conditional Bilingual Mutual Information (CBMI) (Zhang et al., 2022), Self-Evolution (SE) training (Peng et al., 2023)), general machine translation knowledge distillation (Selective Knowledge Distillation (SKD) (Wang et al., 2021), Top-1 Information Enhanced Knowledge Distillation (TIE-KD) (Zhang et al., 2023a)) and machine translation studies for low-frequency word translation (Entropy Regularization (ER) (Pereyra et al., 2017), Linear (Jiang et al., 2019), Exponential and Chi-Square (Gu et al., 2020)).

3.2 Main Results

We compared our approach to BaseFT and other baseline models, especially the previous state-of-the-art method (Gu et al., 2020) on the WMT 16 English→German task. Results on the high/middle/low-frequency test sets are shown in Table 1.

Table 1 shows that: 1) despite previous token weighting and general knowledge distillation studies can significantly improve the overall performance, most of their improvements are on the middle/high-frequency testset and their performances on the low-frequence testset even underperforms the BaseFT baseline, 2) both our method and Gu et al. (2020) do not hamper the performance on the high-frequency test set while improving the performance of both the low-frequency test set and the medium-frequency test set, 3) our single teacher method brings about significantly higher BLEU scores (+0.64) than Gu et al. (2020) on the low-frequency test set and slightly better BLEU scores on both the medium and the high-frequency test sets, and 4) the performance of our dual teacher model only leads to slightly higher BLEU scores than the single teacher model on this task, obtaining +1.24 BLEU improvements on the low-frequency test set compared to BaseFT. But as shown in following experiments (Section 3.3), dual-teacher knowledge distillation is crucial to maintain the performance on the high-frequency test set in more challenging settings with larger training sets than the WMT 16 English→German task.

3.3 Verification on the other Tasks

To validate the effectiveness of our approach on various settings, we conducted experiments on IWSLT 14 German→English, WMT 16 English→German, WMT 15 English→Czech, WMT 18 Chinese→English and WMT 14 English→French tasks, with training set sizes ranging from $174k$ sentence pairs to $36M$ sentence pairs, covering low-resource, middle-resource and high-resource cases. Results are shown in Table 2.

Table 2 shows that: 1) both single-teacher and dual-teacher methods can improve the performance on the low-frequency test sets on all tasks regardless of the training set size, demonstrating the effectiveness of knowledge distillation in improving the performance of low-frequency word translation, 2) the single teacher method can also improve the performance on the middle and high-frequency test sets on low-resource (IWLST 14 German→English) and middle-resource (WMT 16 English→German) tasks compared to the converged base model, obtaining comparable BLEU scores on the high-frequency test set and significantly higher BLEU scores on the middle-frequency test set compared to BaseFT, 3) the performance of the single teacher method on the high-frequency test set decreases significantly on WMT 15 English→Czech, WMT 18 Chinese→English

Task	Model	$Test_{low}$	$Test_{middle}$	$Test_{high}$	$Test_{full}$
IWSLT 14 De->En (174K pairs)	Base	27.66	30.76	35.23	30.97
	BaseFT	27.87	30.88	35.70	31.10
	Low-frequency teacher	28.85	30.72	34.48	30.58
	Single teacher	$28.73^{\dagger}\uparrow$	$31.54^{\dagger}$	$35.67\rightarrow$	31.56
	Dual teacher	$28.74^{\dagger}\uparrow$	$31.59^{\dagger}$	$35.77\rightarrow$	31.69
WMT 16 EN->De (4.5M pairs)	Base	25.17	26.72	28.56	27.15
	BaseFT	25.55	26.97	28.88	27.21
	Low-frequency teacher	26.78	27.43	27.69	27.02
	Single teacher	$26.71^{\dagger}\uparrow$	$27.40^{\dagger}$	$28.97\rightarrow$	27.88
	Dual teacher	$26.79^{\dagger}\uparrow$	$27.44^{\dagger}$	$28.99\rightarrow$	27.93
WMT 15 En->Cs (10M pairs)	Base	27.74	27.98	30.39	28.48
	BaseFT	27.80	28.42	31.40	29.07
	Low-frequency teacher	28.82	28.08	29.09	28.64
	Single teacher	$28.58^{\dagger}\uparrow$	28.06	$29.59\downarrow$	28.68
	Dual teacher	$28.27^{\dagger}\uparrow$	28.64	$31.49\rightarrow$	29.33
WMT 18 Zh->En (19M pairs)	Base	20.95	22.29	25.76	23.22
	BaseFT	21.51	22.48	26.45	23.88
	Low-frequency teacher	22.50	22.62	23.71	22.92
	Single teacher	$22.28^{\dagger}\uparrow$	22.32	$24.71\downarrow$	23.16
	Dual teacher	$22.37^{\dagger}\uparrow$	$22.90^{\ddagger}$	$26.67\rightarrow$	24.05
WMT 14 En->Fr (35M pairs)	Base	37.16	39.58	41.47	39.65
	BaseFT	37.75	40.16	41.93	40.23
	Low-frequency teacher	38.67	40.21	40.22	39.76
	Single teacher	$38.48^{\dagger}\uparrow$	39.37	$40.74\downarrow$	39.26
	Dual teacher	$38.62^{\dagger}\uparrow$	$40.45^{\ddagger}$	$42.01\rightarrow$	40.65

Table 2: Results of single-teacher and dual-teacher methods with increasing training set size. † and ‡ indicate $p < 0.01$ and $p < 0.05$ respectively in the significance test compared to BaseFT.

and WMT 14 English→French tasks with increasing training set sizes, for many cases the single teacher method even under-performs the converged base model, while the dual teacher framework can effectively address this issue and maintain comparable performance on the high-frequency test set compared to BaseFT while obtaining stable improvements on the low-frequency test sets on these challenging tasks.

3.4 Effects of Hyper-parameters

We investigate the effects of the weights of machine translation and distillation losses on performance in Equation 3 on the WMT 16 English→German task. Following Gu et al. (2020), we set the weight of the translation loss (α) to 1 to ensure the learning of the translation task. As for the weight of the low-frequency knowledge distillation loss (γ) and the high-frequency knowledge distillation loss (β), we experimented γ values ranging from 0.3 to 0.7 with an interval of 0.1, and used $1-\gamma$ as corresponding β values. Results are shown in Table 3.

Table 3 shows that: 1) increasing the weight of the low-frequency knowledge distillation loss (γ) consistently improves the performance on the low-frequency test set, but at the cost of the performance on the high-frequency test set, with the performance on the middle-frequency test set improves first and then degrades, and 2) a comparably wide range of choices can ensure the performance on the high-frequency test set and all tested values lead to better performance than BaseFT on the low-frequency test set. We set α, β, and γ to 1, 0.4, and 0.6 respectively for the other experiments as they lead to the best performance on average.

3.5 Effects of Knowledge Distillation Loss Functions

We conducted experiments on the WMT 16 English→German task to test the effects of different knowledge distillation loss functions with the dual-teacher framework. Results are shown in

	γ	β	Dev_{low}	Dev_{middle}	Dev_{high}	Dev_{avg}	$Test_{low}$	$Test_{middle}$	$Test_{high}$	$Test_{avg}$
BaseFT	-	-	23.31	25.03	27.71	25.35	25.55	26.97	28.88	27.13
	0.3	0.7	24.15	25.42	27.89	25.82	26.31	27.20	29.19	27.56
	0.4	0.6	24.42	25.32	27.80	25.85	26.63	27.13	29.16	27.64
Ours	0.5	0.5	24.46	25.22	27.81	25.83	26.65	27.27	29.05	27.65
	0.6	0.4	24.55	25.28	27.77	25.87	26.79	27.44	28.99	27.74
	0.7	0.3	24.68	25.19	27.56	25.81	26.82	27.32	28.69	27.61

Table 3: Results on the WMT 16 English→German task with different hyper-parameters.

	EN->DE		
	$Test_{low}$	$Test_{middle}$	$Test_{high}$
BaseFT	25.55	26.97	28.88
KL Div	26.34	27.04	28.47
Pearson	**26.79**	27.44	28.99

Table 4: Results with different distillation loss functions on the WMT 16 English→German task.

Table 4.

Table 4 shows that: 1) knowledge distillation with both KL divergence and pearson correlation can improve the performance on the low-resource test set, 2) knowledge distillation with pearson correlation leads to more improvements on all test sets than with KL divergence, and 3) the performance on the high-resource test set is worse than BaseFT when distill with the KL divergence loss even with the dual-teacher framework, while knowledge distillation with pearson correlation can lead to slightly higher BLEU scores on the high-frequency test set compared to BaseFT, showing the advantages of knowledge distillation with relative rank than absolute values.

3.6 Case Study

Table 5 shows three translation examples in the IWSLT 14 German→English translation task. In the first sentence, the BaseFT model failed to generate the less frequent noun “stuff” (frequency:951), but used a high-frequency but less proper word “something” (frequency:4235). In the sencond sentence, our method generated the formal form of the less frequent adjective ‘liturgical’ (frequency:103), while the BaseFT model used a more frequent but incorrect word “liturgic” (frequency:675). In the third sentence, our method generate the less frequent but more proper words “favorite” (frequency:265) and “watch” (frequency:446), while the BaseFT model used more frequent but less accurate words “best” (frequency:1094) and “look” (frequency:3889). These examples can be part of the evidence to show the effectiveness of our method.

4 Related Work

4.1 Low-frequency Word Translation

In translation tasks, common types of low-frequency words include rare words, special slang, and technical terminology, among others. The inclusion of low-frequency words in the model’s vocabulary adds diversity but also imposes a significant computational burden on the model. Translation models have limitations when dealing with a large vocabulary. Luong et al. (2015); Jean et al. (2015); Li et al. (2016) attempt to maintain phrase tables or fallback words to address the issue of a large vocabulary. The current mainstream technique involves the use of subword-based methods (Sennrich et al., 2016; Luong and Manning, 2016; Wu et al., 2016), which greatly reduces the vocabulary size and effectively addresses the challenge of representing rare words. Machine translation is essentially a classification task, and there are two main approaches to address the problem of class imbalance: data-based methods (Baloch and Rafi, 2015; Sutskever et al., 2014) and algorithm-based methods (Zhou and Liu, 2005; Lin et al., 2017). Data-based methods primarily employ oversampling and undersampling techniques to address class imbalance. Algorithm-based methods, on the other hand, assign different training strategies to different words. Jiang et al. (2019) propose a linear weighting approach that assigns different weights to words in the translation task based on their frequency, thereby addressing the issue of insufficient translation for low-frequency words. Building upon this, Gu et al. (2020) further introduce chi-square distribution function and power function for weighting, optimizing the translation quality of low-frequency words, achieving the state-of-the-art performance on low-frequency word translation.

Source	zwei Frauen , die existieren und miteinander reden , über irgendetwas .
BaseFT	two women that exist and talk to each other about something .
Ours	two women who exist and talk to each other about stuff .
Reference	two women who exist and talk to each other about stuff .
Source	ihre Arbeit , so denke ich , ist irgendwie liturgisch .
BaseFT	their work , I think , is kind of liturgic .
Ours	their work , I think , is kind of liturgical .
Reference	their work , I think , is kind of liturgical .
Source	wissen Sie , das Beste am Vatersein sind für mich die Filme , die ich schauen kann .
BaseFT	you know , the best thing about father is for me , the films I can look .
Ours	you know , my favorite part of being a dad is the movies I can watch .
Reference	you know , my favorite part of being a dad is the movies I get to watch .

Table 5: Example translations of the BaseFT model and our method.

4.2 Knowledge Distillation

Knowledge distillation is a popular method in recent years to facilitate various transfer learning tasks. Zhuang and Tu (2023) transfer bidirectional language knowledge from masked language pre-training to NMT models. Zhang et al. (2023b) validate that knowledge can be extracted from pre-trained translation models and transferred to student models using knowledge distillation methods. However, a stronger teacher model may not always be beneficial for knowledge distillation, as a significant disparity between the teacher model and the student model may harm the overall performance (Wang et al., 2021). To address this issue, Huang et al. (2022) preserve the relations between the predictions of teacher and student, and propose a correlation-based loss to capture the intrinsic inter-class relations from the teacher explicitly. The problem of class imbalance can be observed in various tasks (Wei et al., 2013; Johnson and Khoshgoftaar, 2019). In the field of image classification, Li et al. (2021) use knowledge distillation techniques to improve imbalanced long-tailed visual recognition tasks. In this paper, we employ knowledge distillation to transfer low-frequency word translation knowledge from the teacher model, aiming at solving the problems brought by the imbalanced word distribution, and present a dual-teacher knowledge distillation framework to preserve the performance on high-frequency words during knowledge distillation.

5 Conclusion

In this study, we investigate the low-frequency word translation problem, which may make the NMT model neglect low-frequency tokens carrying critical semantic information and affect the translation quality. We leverage knowledge distillation to transfer low-frequency word translation knowledge from low-frequency teacher model to the student model. We also present a dual-teacher knowledge distillation framework to ensure the performance with high-frequency words in challenging settings with very large training sets.

Experiment results show that our single-teacher distillation method can already obtain $+0.64$ BLEU improvements over the state-of-the-art method on the WMT 16 English$\rightarrow$German translation task on the low-frequency test set without hampering the performance on the high-frequency test set. While our dual-teacher framework leads to $+0.87$, $+1.24$, $+0.47$, $+0.87$ and $+0.86$ BLEU improvements on the IWSLT 14 German$\rightarrow$English, WMT 16 English$\rightarrow$German, WMT 15 English$\rightarrow$Czech, WMT 14 English$\rightarrow$French and WMT 18 Chinese$\rightarrow$English tasks respectively compared to the fine-tuned baseline, while maintaining the performance on the high-frequency test set even on very large datasets. These results prove the effectiveness of our approach even in very challenging settings.

Limitations

We only tested a number of settings for hyper-parameter selection. But the current setting already shows the effectiveness of our approach, and it is not among the main concern of our work despite that more carefully tuning these hyper-parameters may lead to better performance.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. Yifan Guo and Hongfei Xu acknowledge the support of the National Natural Science Foundation of China (Grant No. 62306284), China Postdoctoral Science Foundation (Grant No. 2023M743189), and the Natural Science Foundation of Henan Province (Grant No. 232300421386). Yifan Guo and Hongying Zan are supported by the National Natural Science Foundation of China (U23A20316).

References

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046.

Çaglar Gulçehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2022. Knowledge distillation from a stronger teacher. In *Advances in Neural Information Processing Systems*, volume 35, pages 33716–33727. Curran Associates, Inc.

Sebastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *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-IJCNLP 2015*, pages 1–10. Association for Computational Linguistics (ACL).

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*, pages 2879–2885.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

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.

Tianhao Li, Limin Wang, and Gangshan Wu. 2021. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 630–639.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *IJCAI*, pages 2852–2858.

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. 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 (Volume 1: Long Papers)*, pages 1054–1063.

Minh-Thang Luong, Ilya Sutskever, Quoc 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 (Volume 1: Long Papers)*, pages 11–19.

Keqin Peng, Liang Ding, Qihuang Zhong, Yuanxin Ouyang, Wenge Rong, Zhang Xiong, and Dacheng Tao. 2023. Token-level self-evolution training for sequence-to-sequence learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 841–850, Toronto, Canada. Association for Computational Linguistics.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions.

Ngoc-Quan Pham, Jan Niehues, and Alex 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*, pages 100–109.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics (ACL).

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yu Wan, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020. Self-paced learning for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1074–1080, Online. Association for Computational Linguistics.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6456–6466, Online. Association for Computational Linguistics.

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*, 16:449–475.

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*.

Yangyifan Xu, Yijin Liu, Fandong Meng, Jiajun Zhang, Jinan Xu, and Jie Zhou. 2021. Bilingual mutual information based adaptive training for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 511–516, Online. Association for Computational Linguistics.

Songming Zhang, Yunlong Liang, Shuaibo Wang, Yufeng Chen, Wenjuan Han, Jian Liu, and Jinan Xu. 2023a. Towards understanding and improving knowledge distillation for neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8062–8079, Toronto, Canada. Association for Computational Linguistics.

Songming Zhang, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, Jian Liu, and Jie Zhou. 2022. Conditional bilingual mutual information based adaptive training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2377–2389.

Yuanchi Zhang, Peng Li, Maosong Sun, and Yang Liu. 2023b. Continual knowledge distillation for neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7996, Toronto, Canada. Association for Computational Linguistics.

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.

Yimeng Zhuang and Mei Tu. 2023. Pretrained bidirectional distillation for machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1132–1145, Toronto, Canada. Association for Computational Linguistics.