

Selective Knowledge Distillation for Neural Machine Translation

Fusheng Wang^{*†1}, Jianhao Yan^{*2}, Fandong Meng², Jie Zhou²

Peking University, China¹

Pattern Recognition Center, WeChat AI, Tencent, China²

wfs0315@pku.edu.com

{elliottyan, fandongmeng}@tencent.com

Abstract

Neural Machine Translation (NMT) models achieve state-of-the-art performance on many translation benchmarks. As an active research field in NMT, knowledge distillation is widely applied to enhance the model’s performance by transferring teacher model’s knowledge on each training sample. However, previous work rarely discusses the different impacts and connections among these samples, which serve as the medium for transferring teacher knowledge. In this paper, we design a novel protocol that can effectively analyze the different impacts of samples by comparing various samples’ partitions. Based on above protocol, we conduct extensive experiments and find that the teacher’s knowledge is not the more, the better. Knowledge over specific samples may even hurt the whole performance of knowledge distillation. Finally, to address these issues, we propose two simple yet effective strategies, i.e., batch-level and global-level selections, to pick suitable samples for distillation. We evaluate our approaches on two large-scale machine translation tasks, WMT’14 English-German and WMT’19 Chinese-English. Experimental results show that our approaches yield up to +1.28 and +0.89 BLEU points improvements over the Transformer baseline, respectively. ¹

1 Introduction

Machine translation has made great progress recently by using sequence-to-sequence models (Sutskever et al., 2014; Vaswani et al., 2017; Meng and Zhang, 2019; Zhang et al., 2019b; Yan et al., 2020). Recently, some knowledge distillation methods (Kim and Rush, 2016; Freitag et al., 2017; Gu

et al., 2017; Tan et al., 2019; Wei et al., 2019; Li et al., 2020; Wu et al., 2020) are proposed in the machine translation to help improve model performance by transferring knowledge from a teacher model. These methods can be divided into two categories: word-level and sequence-level, by the granularity of teacher information. In their researches, the model learns from teacher models by minimizing gaps between their outputs on every training word/sentence (i.e., corresponding training sample) without distinction.

Despite their promising results, previous studies mainly focus on finding what to teach and rarely investigate how these words/sentences (i.e., samples), which serve as the medium or carrier for transferring teacher knowledge, participate in the knowledge distillation. Several questions remain unsolved for these samples: Which part of all samples shows more impact in knowledge distillation? Intuitively, we may regard that longer sentences are hard to translate and might carry more teacher knowledge. But are there more of these criteria that can identify these more important/suitable samples for distillation? Further, what are the connections among these samples? Are they all guiding the student model to the same direction? By investigating the carrier of teacher knowledge, we can shed light on finding the most effective KD method.

Hence, in this paper, we aim to investigate the impacts and differences among all samples. However, it is non-trivial to analyze each of them. Therefore, we propose a novel analytical protocol by partitioning the samples into two halves with a specific criterion (e.g., sentence length or word cross-entropy) and study the gap between performance. Extensive empirical experiments are conducted to analyze the most suitable sample for transferring knowledge. We find that different samples differ in transferring knowledge for a substantial margin. More interestingly, with some partitions, especially the student

^{*}Equal contribution.

[†]This work was done when Fusheng Wang was interning at Pattern Recognition Center, Wechat AI, Tencent Inc, China.

¹We release our code on <https://github.com/LeslieOverfitting/selective.distillation>.

model’s word cross-entropy, the model with half of the knowledge even shows better performance than the model using all distill knowledge. The benefit of the distillation of two halves cannot collaborate. This phenomenon reveals that the distillation of two halves cannot collaborate, even hurt the whole performance. Hence, a more sophisticated selective strategy is necessary for KD methods.

Next, we propose two simple yet effective methods to address the observed phenomenon according to word cross-entropy (Word CE), which we find is the most distinguishable criterion. We first propose a batch-level selection strategy that chooses words with higher Word CE within the current batch’s distribution. Further, to step forward from local (batch) distribution to global distribution, we use a global-level FIFO queue to approximate the optimal global selection strategy, which caches the Word CE distributions across several steps. We evaluate our proposed method on two large-scale machine translation datasets: WMT’14 English-German and WMT’19 Chinese-English. Experimental results show that our approach yields an improvement of +1.28 and + 0.89 BLEU points over the Transformer baseline.

In summary, our contributions are as follows:

- We propose a novel protocol for analyzing the property for the suitable medium samples for transferring teacher’s knowledge.
- We conduct extensive analyses and find that some of the teacher’s knowledge will hurt the whole effect of knowledge distillation.
- We propose two selective strategies: batch-level selection and global-level selection. The experimental results validate our methods are effective.

2 Related Work

Knowledge distillation approach (Hinton et al., 2015) aims to transfer knowledge from teacher model to student model. Recently, many knowledge distillation methods (Kim and Rush, 2016; Hu et al., 2018; Sun et al., 2019; Tang et al., 2019; Jiao et al., 2019; Zhang et al., 2019a, 2020; Chen et al., 2020a; Meng et al., 2020) have been used to get effective student model in the field of natural language processing by using teacher model’s outputs or hidden states as knowledge.

As for neural machine translation (NMT), knowledge distillation methods commonly focus on bet-

ter improving the student model and learning from the teacher model. Kim and Rush (2016) first applied knowledge distillation to NMT and proposed the sequence-level knowledge distillation that lets student model mimic the sequence distribution generated by the teacher model. It was explained as a kind of data augmentation and regularization by Gordon and Duh (2019). Further, Freitag et al. (2017) improved the quality of distillation information by using an ensemble model as the teacher model. Gu et al. (2017) improved non-autoregressive model performance by learning distillation information from the autoregressive model. Wu et al. (2020) proposed a layer-wise distillation method to be suitable for the deep neural network. Chen et al. (2020b) let translation model learn from language model to help the generation of machine translation.

To the best of our knowledge, there is no previous work in NMT concerning the selection of suitable samples for distillation. The few related ones mainly focus on selecting appropriate teachers for the student model to learn. For instance, Tan et al. (2019) let the student model only learn from the individual teacher model whose performance surpasses it. Wei et al. (2019) proposed an online knowledge distillation method that let the model selectively learn from history checkpoints. Unlike the above approaches, we explore the effective selective distillation strategy from sample perspective and let each sample determine learning content and degree.

3 Background

3.1 Neural Machine Translation

Given a source sentence $\mathbf{x} = (x_1, \dots, x_n)$, and its corresponding ground-truth translation sentence $\mathbf{y} = (y_1^*, \dots, y_m^*)$, an NMT model minimizes the word negative log-likelihood loss at each position by computing cross-entropy. For the j -th word in the target sentence, the loss can be formulated as:

$$\mathcal{L}_{ce} = - \sum_{k=1}^{|V|} \mathbb{1}\{y_j^* = k\} \log p(y_j = k | \mathbf{y}_{<j}, \mathbf{x}; \theta), \quad (1)$$

where $|V|$ is the size of target vocabulary, $\mathbb{1}$ is the indicator function, and $p(\cdot|\cdot)$ denotes conditional probability with model parameterized by θ .

Criteria	BLEU		
	\mathcal{S}_{High}	\mathcal{S}_{Low}	Δ
Baseline	27.29		-
Distill-All	28.14		-
Distill-Half(Random)	28.18		-
Data Property			
Sentence Length	27.81	27.59	+0.22
Word Frequency	28.35	27.99	+0.36*
Student Model			
Embedding Norm	27.90	27.73	+0.17
Word CE	28.42	27.78	+0.64*
Sentence CE	28.29	27.84	+0.45*
Teacher Model			
Teacher P_{golden}	27.97	28.00	-0.03
Entropy	27.62	27.92	-0.30

Table 1: BLEU score (%) of different criteria in WMT’14 En-De. Δ denotes the difference of BLEU score (%) between \mathcal{S}_{High} and \mathcal{S}_{Low} . “*”: significantly ($p < 0.05$) difference between the \mathcal{S}_{High} and \mathcal{S}_{Low} .

3.2 Word-level Knowledge Distillation

In knowledge distillation, student model S gets extra supervision signal by matching its own outputs to the probability outputs of teacher model T . Specifically, word-level knowledge distillation defines the Kullback–Leibler distance between the output distributions of student and teacher (Hu et al., 2018). After removing constants, the objective is formulated as:

$$\mathcal{L}_{kd} = - \sum_{k=1}^{|V|} q(y_j = k | \mathbf{y}_{<j}, \mathbf{x}; \theta_T) \times \log p(y_j = k | \mathbf{y}_{<j}, \mathbf{x}; \theta_S), \quad (2)$$

where $q(\cdot|\cdot)$ is the conditional probability of teacher model. θ_S and θ_T is the parameter set of student model and teacher model, respectively.

And then, the overall training procedure is minimizing the summation of two objectives:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{kd}, \quad (3)$$

where α is a weight to balance two losses.

4 Are All Words Equally Suitable for KD?

As discussed before, as a carrier of the teacher’s knowledge, ground-truth words might greatly influence the performance of knowledge distillation. Therefore, in this section, we first do some preliminary empirical studies to evaluate the importance

of different words/sentences in knowledge distillation.

4.1 Partition of Different Parts

The optimal way to analyze samples’ different impacts on distillation is to do ablation studies over each of them. However, it is clearly time-consuming and intractable. Hence, we propose an analytical protocol by using the partition and comparison as an approximation, which we believe could shed light on future analyses. Particularly, we leverage a specific criterion f to partition samples into two complementary parts:

$$\mathcal{S}_{High} := \{y_i \mid f(y_i) > \text{Median}(f(\mathbf{y})), y_i \in \mathbf{y}\},$$

$$\mathcal{S}_{Low} := \{y_i \mid f(y_i) \leq \text{Median}(f(\mathbf{y})), y_i \in \mathbf{y}\},$$

and analyze different effects between \mathcal{S}_{High} and \mathcal{S}_{Low} . Each part consists of 50% words/sentences precisely. The criteria come from three different perspectives: data property, student model, and teacher model. The detailed descriptions are as follows:

- **Data Property.**

As longer sentences and rare words are more challenging to translate (Kocmi and Bojar, 2017; Platanios et al., 2019), its corresponding teacher knowledge may benefit the student model more. Hence, we choose sentence length and word frequency as criteria.

- **Student Model.** As for the student model, we care if the student model thinks these words/sentences are too complicated. Therefore, we use Word CE (cross-entropy of words), Sentence CE (mean of the cross-entropy of all words in sentences), and each word’s embedding norm (Liu et al., 2020).

- **Teacher Model.** For the teacher model, we guess that the teacher’s prediction confidence may be crucial for transferring knowledge. Hence, we use the prediction probability of ground-truth label (P_{golden}) and entropy of prediction distribution as our criteria.

4.2 Analytic Results

Table 1 presents our results on different criteria. We also add the performance of Transformer baseline, Distill-All (distillation with all words) and Distill-Half(distillation with 50% words chosen by random) for comparison.

Impact of Different Parts. Through most of the rows, we observe noticeable gaps between the BLEU scores of the \mathcal{S}_{High} and \mathcal{S}_{Low} , indicating there exists a clear difference of impact on medium of teacher knowledge. Specifically, for most of the criteria like cross-entropies or word frequency, the gap between two halves surpasses 0.35. In contrast, teacher P_{golden} seems not useful for partitioning KD knowledge. We conjecture this is because no matter whether the teacher is convinced with the golden label or not, other soft labels could contain useful information (Gou et al., 2020). Besides, we find teacher entropy is a good-enough criterion for partitioning KD data, which inlines with previous studies of dark knowledge (Dong et al., 2019). Finally, we find that the KD is most sensitive (+0.64) with the Word CE criterion, which enjoys the adaptivity during the training phase and is a good representative for whether the student thinks the sample is difficult.

In conclusion, we regard the most suitable samples should have the following properties: higher Word CE, higher Sentence CE, higher Word Frequency, which probably benefits future studies of effective KD methods.

Impact of All and Halves. More interestingly, compared with ‘Distill-All’, which is the combination of the \mathcal{S}_{High} and \mathcal{S}_{Low} , the \mathcal{S}_{High} halves’ BLEU score even surpass the ‘Distill-All’, for Word CE, Sentence CE and Word Frequency criteria. This leads to two conclusions:

- (1) Within some partitions, the \mathcal{S}_{High} contributes most to the KD improvements.
- (2) The amount of teacher knowledge is not the more, the better. The distillation knowledge of the \mathcal{S}_{Low} does not directly combine with the \mathcal{S}_{High} , even hurts \mathcal{S}_{High} ’s performance.

Impact of the Amount of Knowledge. Given that distillation knowledge is most sensitive to Word CE, we conduct extra analysis on the Word CE. Figure 1 presents the results of varying the amount of knowledge for \mathcal{S}_{High} and \mathcal{S}_{Low} . The consistent phenomenon is that the \mathcal{S}_{High} perform significantly better than the \mathcal{S}_{Low} when using the same amount of teacher’s knowledge. These results suggest that we should focus more on the \mathcal{S}_{High} than on \mathcal{S}_{Low} . Besides, we notice that the model performance increases when we increase the knowledge in \mathcal{S}_{High} , but not the case for \mathcal{S}_{Low} . We conclude that the Word CE is distinguishable and a

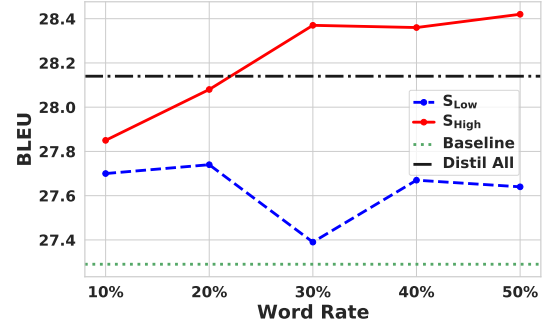


Figure 1: BLEU score (%) on WMT’14 En-De translation task. \mathcal{S}_{Low} means the subset of training set which have relative small word-level cross-entropy and easy for model to learn. \mathcal{S}_{High} means the subset of training set which have relative large word-level cross-entropy and hard for model to learn. ‘Word Rate’ controls the number of words need to get extra distillation knowledge from teacher model. For example, word rate=30% means that student model only learns distillation knowledge of words whose cross-entropy loss in biggest / smallest 30%. We choose the model which performs the best on the validation set and report its performance on test sets.

better indicator of teachers’ useful knowledge only for \mathcal{S}_{High} .

At the end of this section, we can summary the following points:

- To find out the most suitable medium for transferring medium, we adopt a novel method of partition and comparison, which can easily be adopted to future studies.
- The benefit of distillation knowledge drastically changes when applying to different mediums of knowledge.
- Among all criteria, knowledge distillation is the most sensitive to Word CE. Distilling words with higher Word CE is more reliable than words with lower CE.
- In some partitions, the distillation benefit of \mathcal{S}_{Low} can not add to the \mathcal{S}_{High} , even hurts \mathcal{S}_{High} ’s performance.

5 Selective Knowledge Distillation for NMT

As mentioned above, there exist un-suitable mediums/samples that hurt the performance of knowledge distillation. In this section, we address this problem by using two simple yet effective strategy of selecting useful samples.

In Section 4, we find that Word CE is the most distinguishable criterion. Hence, we continue to

use the Word CE as the measure in our methods. As the word cross-entropy is a direct measure of how the student model agrees with the golden label, we refer to words with relatively large cross-entropy as difficult words, and words with relatively small cross-entropy as easy words, in the following parts. This is to keep the notation different from previous analysis.

Then, we only need to define what is “relatively large”. Here, we introduce two CE-based selective strategies:

Batch-level Selection (BLS). Given a mini-batch B of sentence pairs with M target words, we sort all words in the current batch with their Word CE in descending order and select the top r percent of all words to distill teacher knowledge. More formally, let \mathcal{A} denote the Word CE set, which contains the Word CE of each word in batch B . We define $\mathcal{S}_{Hard} = \text{top-}r\%(\mathcal{A})$ as the set of the $r\%$ largest cross-entropy words among the batch, and \mathcal{S}_{Easy} is its complementary part.

For those words in \mathcal{S}_{Hard} , we let them get extra supervision signal from teacher model’s distillation information. Therefore, the knowledge distillation objective in Equation 3 can be re-formulated as:

$$\mathcal{L}_{kd} = \begin{cases} -\sum_{k=1}^{|V|} q(y_k) \cdot \log p(y_k), & y \in \mathcal{S}_{Hard} \\ 0, & y \in \mathcal{S}_{Easy} \end{cases}$$

where we simplify the notation of p and q for clarity.

Global-level Selection (GLS). Limited by the number of words in a mini-batch, batch-level selection only reflects the current batch’s CE distribution and can not represent the real global CE distribution of the model very well. In addition, the batch-level method makes our relative difficulty measure easily affected by each local batch’s composition. The optimal approach to get the global CE distribution is to traverse all training set words and calculate their CE to get the real-time distribution after each model update. However, this brings a formidable computational cost and is not realistic in training.

Therefore, as a proxy to optimal way, we extend batch-level selection to global-level selection by dexterously using a First-In-First-Out (FIFO) global queue \mathcal{Q} . At each training step, we push batch words’ CE into FIFO global queue \mathcal{Q} and pop out the ‘Oldest’ words’ CE in the queue to retain the queue’s size. Then, we sort all CE values in the queue and calculate the ranking position

Algorithm 1 Global-level Selection

Input: B : mini-batch, \mathcal{Q} : FIFO global queue, \mathcal{T} : teacher model, \mathcal{S} : student model

- 1: **for** each $word_i$ in B **do**
 - 2: Compute \mathcal{L}_{ce} of $word_i$ by Equation 1
 - 3: Compute \mathcal{L}_{kd} of $word_i$ by Equation 2
 - 4: Push \mathcal{L}_{ce} to \mathcal{Q}
 - 5: **if** L_{ce} in $\text{top-}r\%(\mathcal{Q})$ **then**
 - 6: $Loss_i \leftarrow \mathcal{L}_{ce} + \alpha \cdot \mathcal{L}_{kd}$
 - 7: **else**
 - 8: $Loss_i \leftarrow \mathcal{L}_{ce}$
 - 9: $Loss \leftarrow Loss + Loss_i$
 - 10: Update \mathcal{S} with respect to $Loss$
-

of each word. The storage of queue is much bigger than a mini-batch so that we can evaluate the current batch’s CEs with more words, which reduces the fluctuation of CE distribution caused by the batch-level one. Algorithm 1 details the entire procedure.

6 Experiments

We carry out experiments on two large-scale machine translation tasks: WMT’14 English-German (En-De) and WMT’19 Chinese-English (Zh-En).

6.1 Setup

Datasets. For WMT’14 En-De task, we use 4.5M preprocessed data, which is tokenized and split using byte pair encoded (BPE) (Sennrich et al., 2016) with 32K merge operations and a shared vocabulary for English and German. We use *newstest2013* as the validation set and *newstest2014* as the test set, which contain 3000 and 3003 sentences, respectively.

For the WMT’19 Zh-En task, we use 20.4M preprocessed data, which is tokenized and split using 47K/32K BPE merge operations for source and target languages. We use *newstest2018* as our validation set and *newstest2019* as our test set, which contain 3981 and 2000 sentences, respectively.

Evaluation. For evaluation, we train all the models with a maximum of 300K steps for WMT En-De’14 and WMT’19 Zh-En. We choose the model which performs the best on the validation set and report its performance on test set. We measure case sensitive BLEU calculated by *multi-bleu.perl*²

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

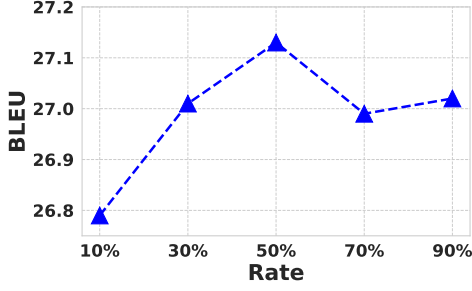


Figure 2: BLEU score (%) with different $r\%$ on validation set of WMT’14 En-De.

and `mteval-v13a.pl`³ with significance test (Koehn, 2004) for WMT’14 En-De and WMT’19 Zh-En, respectively.

Model and Hyper-parameters. Following the setting in Vaswani et al. (2017), we carry out our experiments on standard Transformer (Vaswani et al., 2017) with the fairseq toolkit (Ott et al., 2019). By default, we use Transformer (Base), which contains six stacked encoder layers and six stacked decoder layers as both teacher model and student model. To verify our approaches can be applied to a stronger teacher and student models, we further use deep Transformers with twelve encoder layers and six decoder layers. In training processing, we use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, learning rate is $7e-4$ and dropout is 0.1. All experiments are conducted using 4 NVIDIA P40 GPUs, where the batch size of each GPUs is set to 4096 tokens. And we accumulate the gradient of parameters and update every two steps. The average runtimes are 3 GPU days for all experiments.

There are two hyper-parameters in our experiment, i.e., distil rate $r\%$ and global queue size Q_{size} . For distil rate $r\%$, the search space is [10%, 30%, 50%, 70%, 90%]. The search result of $r\%$ is shown in Figure 2, we can find that the performance is sensitive to the value of $r\%$. When the ratio is smaller than 50%, the increase of ratio is consistent with the BLEU score increases, and the best performance peaks at 50%. We directly apply the distil rate $r\%$ to the WMT’19 Zh-En task without extra searching. Besides, We set the $Q_{size} = 30K$ for WMT’14 En-De. For larger dataset WMT’19 Zh-En, we enlarge the Q_{size} to from 30K to 50K and keep word rate unchanged. The hyper-parameter search of Q_{size} can be found in Section 6.4.

³<https://github.com/moses-sm/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl>

Models	En-De	Δ
<i>Existing NMT systems</i>		
Vaswani et al. (2017)	27.30	ref
Vaswani et al. (2017) (Big)	28.40	+1.10
Chen et al. (2020b)	27.53	+0.23
Zheng et al. (2019)	28.10	+0.80
So et al. (2019)	28.40	+1.10
Tay et al. (2020)	28.47	+1.17
<i>Our Implemented Methods</i>		
Transformer	27.29	ref
Word-KD	28.14	+0.85
Seq-KD	28.15	+0.86
Batch-level Selection	28.42*	+1.13
Global-level Selection	28.57*†	+1.28

Table 2: BLEU scores (%) on WMT’14 English-German (En-De) task. Δ shows the improvement compared to Transformer (Base). ‘*’: significantly ($p < 0.01$) better than Transformer (Base). ‘†’: significantly ($p < 0.05$) better than the Word/Seq-KD models.

Compared Methods. We compare our method with several existing NMT systems (KD and others):

- **Word-KD (Kim and Rush, 2016).** Word-KD is a standard method that distills knowledge equally for each word. The detailed description is in Section 3.2.
- **Seq-KD (Kim and Rush, 2016).** Sequence-KD uses teacher generated outputs on training corpus as an extra source. The training loss can be formulated as:

$$\mathcal{L}_{seq_kd} = - \sum_{j=1}^J \sum_{k=1}^{|V|} \mathbb{1}\{\hat{y}_j = k\} \times \log p(y_j = k | \hat{\mathbf{y}}_{<j}, \mathbf{x}; \theta), \quad (4)$$

where $\hat{\mathbf{y}}$ denotes the sequence predicted by teacher model from running beam search, J is the length of target sentence.

- **Bert-KD (Chen et al., 2020b).** This method leverages the pre-trained Bert as teacher model to help NMT model improve machine translation quality.
- **Other Systems.** We also include some existing methods based on Transformer(Base) for comparison, i.e., Zheng et al. (2019); So et al. (2019); Tay et al. (2020).

6.2 Main Results

Results on WMT’14 English-German. The results on WMT’14 En-De are shown in Table 2. In

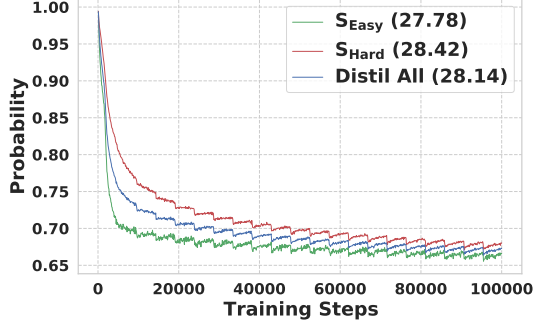


Figure 3: The probability for gradients of \mathcal{L}_{kd} and \mathcal{L}_{ce} pointing the same direction.

this experiment, both the teacher model and student model are Transformer (Base). We also list our implementation of word-level distillation and sequence level distillation (Kim and Rush, 2016) method.

Firstly, compared with the Transformer (Base), our re-implemented word-level and the sequence-level distillation show similar improvements with the BLEU scores up from 27.29 to 28.14 and 28.15, respectively. Secondly, compared with these already strong baseline methods, our batch-level selective approach further extends the improvement to 28.42, proving the selective strategy’s effectiveness. Thirdly, our global-level distillation achieves a 28.57 BLEU score and outperforms all previous methods, showing that the better evaluation of words’ CE distribution with FIFO global queue helps selection. It is worth noting that our strategy also significantly improves translation quality over all others methods including Word-KD. Finally, our methods show comparable/better performance than other existing NMT systems and even surpass the Transformer (Big), with much fewer parameters.

6.3 Analysis

Even though we find some interesting phenomena and achieve great improvement by selective distillation, the reason behind it is still unclear. Hence, in this section, we conduct some experiments to analyze and explain the remaining question.

Note that we follow the previous partition and comparison method in this section and divide the samples with/without KD loss defined in our selection strategy as $\mathcal{S}_{Hard}/\mathcal{S}_{Easy}$.

Conflict on Different Parts. The first question is that why our methods surpass the Word-KD with more knowledge. To answer this question, we collect the statistics on the gradient difference between

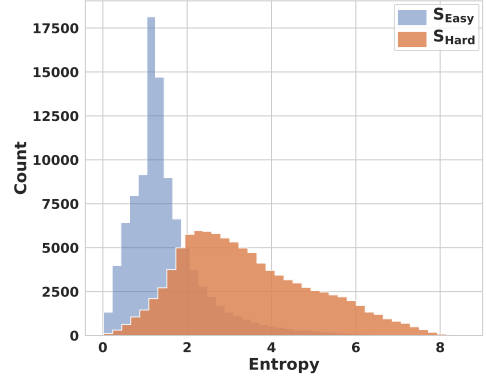


Figure 4: The entropy of prediction distribution of teacher model for different parts.

knowledge distillation loss and cross-entropy loss on the ground-truth label for \mathcal{S}_{Hard} and \mathcal{S}_{Easy} .

Here, we study gradients over the output distributions, which are directly related to the model’s performance. Particularly, decoder maps target sentences $\mathbf{y} = (y_1^*, \dots, y_m^*)$ to their corresponding hidden representation $\mathbf{h} = (h_1, \dots, h_m)$. For words in target sequence, the prediction logits $\mathbf{l} \in \mathbb{R}^{d_{model} \times |V|}$ is given by:

$$\mathbf{l} = \mathbf{h}^T \mathcal{W} \quad (5)$$

$$\mathbf{p} = \text{Softmax}(\mathbf{l}) \quad (6)$$

where $\mathbf{h} \in \mathbb{R}^{d_{model}}$ is the layer output of transformer decoder, $\mathcal{W} \in \mathbb{R}^{d_{model} \times |V|}$ is projection matrix. Then, the gradient respect to \mathbf{l} from golden cross-entropy loss can be denotes as $\nabla_{\mathbf{l}} \mathcal{L}_{ce}$. The gradient from distillation loss can be denotes as $\nabla_{\mathbf{l}} \mathcal{L}_{kd}$. Next, we calculate the probability that $\nabla_{\mathbf{l}} \mathcal{L}_{ce}$ and $\nabla_{\mathbf{l}} \mathcal{L}_{kd}$ share the same direction.

Figure 3 presents the results with the probability that gradients agree with each other during training. We observe that \mathcal{S}_{Easy} (green line) is consistently lower than distillation with all words (blue line) and \mathcal{S}_{Hard} (red line), which means \mathcal{S}_{Easy} has more inconsistency with ground-truth. Combining with the BLEU performances, we argue this consistency leads to the risk of introducing noise and disturbs the direction of parameter updating.

Besides, the agreement of Distill-All (blue line in Fig) lies in the middle of two halves. It proves that \mathcal{S}_{Easy} and \mathcal{S}_{Hard} compromise with each other on some conflicts. It also proves that there exist some conflicts between the knowledge in \mathcal{S}_{Easy} and \mathcal{S}_{Hard} .

Knowledge on Different Parts. In our approaches, we select the transferring samples from

Models	Zh-En	Δ
Transformer (Base)	25.73	ref
Word-KD	26.21	+0.48
Seq-KD	27.27	+1.54
Word-KD + Ours	26.62*	+0.89
Seq-KD + Ours	27.61*	+1.88

Table 3: BLEU scores (%) on WMT’19 Chinese-English (Zh-En) task. Δ shows the improvement compared to Transformer (Base). ‘*’: significantly ($p < 0.01$) better than the Transformer (Base).

the student model’s point of view. However, in previous literature, they commonly consider knowledge from the teacher’s perspective. Hence, in this section, we study the correlation between these two perspectives.

Because previous studies commonly regard teacher’s soft-labels contain dark knowledge (Dong et al., 2019), we take the entropy of teacher’s prediction as a proxy. Concretely, we randomly select 100K tokens in the training set and calculate the entropy of distribution predicted by the teacher model for both \mathcal{S}_{Hard} and \mathcal{S}_{Easy} . As shown in Figure 4, we notice that the \mathcal{S}_{Easy} ’s entropy distribution is more concentrated in range (0, 4) and peaks around 1.2. In contrast, the \mathcal{S}_{Hard} ’s entropy distribution is more spread out. The overall distribution shifts to higher entropy, which indicates \mathcal{S}_{Hard} tends to provide a smoother supervision signal. Consequently, we conclude that even though our selective strategy comes from the student’s perspective, it also favors samples with abundant dark knowledge in teacher’s perspective. To some extent, this explains why the \mathcal{S}_{Hard} ’ knowledge benefits distillation performance more.

6.4 Generalizability

Results on WMT’19 Chinese-English. We also conduct experiments on the larger WMT’19 Zh-en dataset (20.4M sentence pairs) to ensure our methods can provide consistent improvements across different language pairs.

As shown in Table 3, our method still significantly outperforms the Transformer (Base) with +0.89. Compared with the Word-KD, our approach consistently improves with +0.41 BLEU points. Besides, we also find that Seq-KD with our methods extends the improvement of BLEU score from 27.27 to 27.61. This indicates that our selective strategy is partially orthogonal to the improvement

Models	En-De	Δ
Deep Transformer (12 + 6)	27.94	ref
Word-KD	28.90	+0.96
Ours	29.12*	+1.18

Table 4: BLEU scores (%) on WMT’14 English-German (En-De) task. Here we use Deep Transformers (12 encoders and 6 decoders) for both the teacher and student model. Δ shows the improvement compared to Deep Transformer (12 + 6). ‘*’: significantly ($p < 0.01$) better than Deep Transformer (12 + 6).

of Seq-KD and maintains generalizability. In summary, these results suggest that our methods can achieve consistent improvement on different sized datasets across different language pairs.

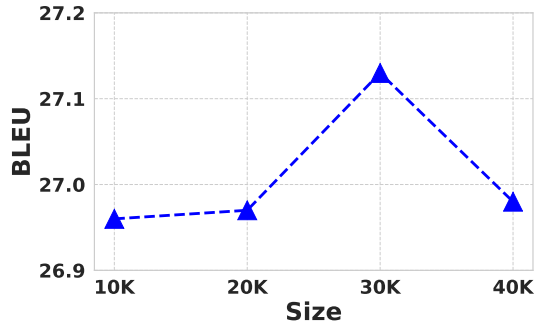
Results with Larger Model Size. Here, we investigate how our method is well-generalized to larger models. We use a deep transformer model with twelve encoder layers and six decoder layers for our larger model experiments. As shown in Table 4, Deep Transformer (12 + 6) and Word-KD have already achieved strong performance with up to 28.90 BLEU points, and our method still outperforms these baselines (29.12 BLEU). It proves our methods’ generalizability to larger models.

6.5 Effect of the Global Queue

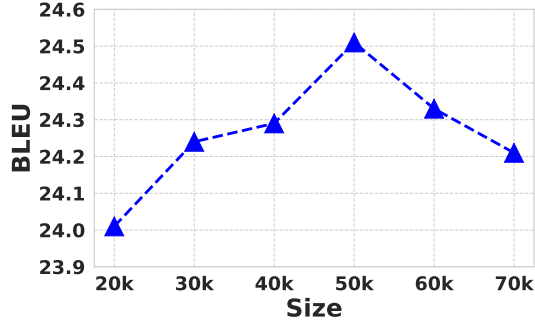
This section analyzes how Q_{size} affects our model’s performance. As mentioned before, Q_{size} denotes the size of the global FIFO queue, which affects simulating the word cross-entropy distribution of the current model.

Figure 5 shows the search results of Q_{size} . We can find that smaller and larger queue size both hurts the BLEU scores. Besides, 30K and 50K of queue size are the best for WMT’14 En-De and WMT’19 Zh-En, respectively. This also accords with our intuition that smaller Q_{size} degrades the global-level queue to batch level, and larger Q_{size} slows down the update of CE distribution.

Figure 6 plots the partition Word CE of \mathcal{S}_{Hard} and \mathcal{S}_{Easy} for batch-level and global-level selection. We can see that, as the training progresses, batch-level selection starts to suffer from the high variance because of each batch’s randomness. Selections with FIFO queue drastically reduce the variance and make a reasonable estimation of global CE distribution. These findings prove the effectiveness of our proposed FIFO queue.



(a) En-De



(b) Zh-En

Figure 5: BLEU score (%) with different Q_{size} on WMT’14 En-De and WMT’19 Zh-En validation set.

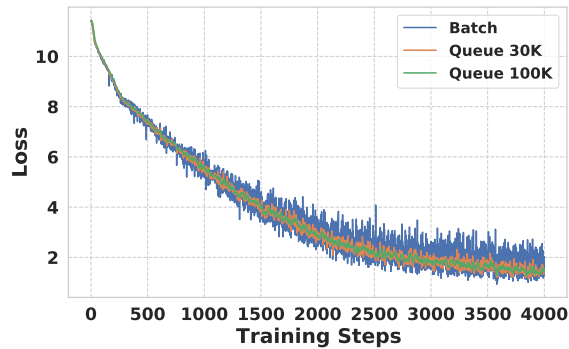


Figure 6: Partition point for \mathcal{S}_{Hard} and \mathcal{S}_{Easy} , with respect to different strategies. Batch-level selection clearly suffers from large fluctuations and high variance.

7 Conclusion

In this work, we conduct an extensive study to analyze the impact of different words/sentences as the carrier in knowledge distillation. Analytic results show that distillation benefits have a substantial margin, and these benefits may not collaborate with their complementary parts and even hurt the performance. To address this problem, we propose two simple yet effective strategies, namely the batch-level selection and global-level selection.

The experiment results show that our approaches can achieve consistent improvements on different sized datasets across different language pairs.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

- Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020a. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437, Online. Association for Computational Linguistics.
- Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020b. Distilling knowledge learned in bert for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7893–7905.
- Bin Dong, Jikai Hou, Yiping Lu, and Zhihua Zhang. 2019. Distillation \approx early stopping? harvesting dark knowledge utilizing anisotropic information retrieval for overparameterized neural network. *arXiv preprint arXiv:1910.01255*.
- Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. 2017. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*.
- Mitchell A Gordon and Kevin Duh. 2019. Explaining sequence-level knowledge distillation as data-augmentation for neural machine translation. *arXiv preprint arXiv:1912.03334*.
- Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. 2020. Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*.
- Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou. 2018. Attention-guided answer distillation for machine reading comprehension. *arXiv preprint arXiv:1808.07644*.

- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.
- Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.
- Tom Kocmi and Ondrej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. *arXiv preprint arXiv:1707.09533*.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.
- Bei Li, Ziyang Wang, Hui Liu, Quan Du, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2020. Learning light-weight translation models from deep transformer. *arXiv preprint arXiv:2012.13866*.
- Xuebo Liu, Houtim Lai, Derek F Wong, and Lidia S Chao. 2020. Norm-based curriculum learning for neural machine translation. *arXiv preprint arXiv:2006.02014*.
- Fandong Meng, Jianhao Yan, Yijin Liu, Yuan Gao, Xianfeng Zeng, Qinsong Zeng, Peng Li, Ming Chen, Jie Zhou, Sifan Liu, and Hao Zhou. 2020. [WeChat neural machine translation systems for WMT20](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 239–247, Online. Association for Computational Linguistics.
- Fandong Meng and Jinchao Zhang. 2019. DTMT: A novel deep transition architecture for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 224–231.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.
- Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M Mitchell. 2019. Competence-based curriculum learning for neural machine translation. *arXiv preprint arXiv:1903.09848*.
- 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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- David So, Quoc Le, and Chen Liang. 2019. The evolved transformer. In *International Conference on Machine Learning*, pages 5877–5886. PMLR.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27:3104–3112.
- Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*.
- Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*.
- Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2020. Synthesizer: Rethinking self-attention in transformer models. *arXiv preprint arXiv:2005.00743*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Hao-Ran Wei, Shujian Huang, R. Wang, Xin-Yu Dai, and Jiajun Chen. 2019. Online distilling from checkpoints for neural machine translation. In *NAACL-HLT*.
- Yimeng Wu, Peyman Passban, Mehdi Rezagholizade, and Qun Liu. 2020. Why skip if you can combine: A simple knowledge distillation technique for intermediate layers. *arXiv preprint arXiv:2010.03034*.
- Jianhao Yan, Fandong Meng, and Jie Zhou. 2020. Multi-unit transformers for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1047–1059, Online.
- Biao Zhang, Deyi Xiong, Jinsong Su, and Jiebo Luo. 2019a. Future-aware knowledge distillation for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2278–2287.
- Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, and Qun Liu. 2020. Ternarybert: Distillation-aware ultra-low bit bert. *arXiv preprint arXiv:2009.12812*.
- Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019b. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy.

Zaixiang Zheng, Shujian Huang, Zhaopeng Tu, Xin-Yu Dai, and Jiajun Chen. 2019. Dynamic past and future for neural machine translation. *arXiv preprint arXiv:1904.09646*.