KNOWLEDGE DISTILLATION BASED ENSEMBLE LEARNING FOR NEURAL MACHINE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Model ensemble can effectively improve the accuracy of neural machine translation, which is accompanied by the cost of large computation and memory requirements. Additionally, model ensemble cannot combine the strengths of translation models with different decoding strategies since their translation probabilities cannot be directly aggregated. In this paper, we introduce an ensemble learning framework based on knowledge distillation to aggregate the knowledge of multiple teacher models into a single student model. Under this framework, we introduce word-level ensemble learning and sequence-level ensemble learning for neural machine translation, where sequence-level ensemble learning is capable of aggregating translation models with different decoding strategies. Experimental results on multiple translation tasks show that, by combining the two ensemble learning methods, our approach achieves substantial improvements over the competitive baseline systems and establishes a BLEU score of 31.13 in the WMT14 English-German translation task.[1]

1 INTRODUCTION

Neural Machine Translation (NMT) has achieved impressive performance over the past several years (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017). Model ensemble is an effective and widely used technique to improve the translation accuracy of NMT. However, model ensemble requires to simultaneously run multiple models during decoding and take their average probabilities for prediction, which significantly increases the computational cost and memory requirements. Another limitation is that it cannot aggregate translation models with different decoding strategies. For example, although it is known that left-to-right (L2R) and right-to-left (R2L) translation models are complementary (Zhang et al., 2018; 2020), these two kinds of models cannot be directly aggregated due to their different translation directions.

Knowledge distillation approaches (Hinton et al., 2015) learn a smaller student network through distilling knowledge from a larger teacher network. It is a natural idea to apply this technique to aggregate the knowledge of multiple models into a single model (Furlanello et al., 2018; Fukuda et al., 2017; Zhu et al., 2018; Asif et al., 2019). In neural machine translation, sequence-level knowledge distillation is proposed by Kim & Rush (2016) and then applied in Ensemble Distillation (Freitag et al., 2017), Online Distillation (Wei et al., 2019) and Transductive Ensemble Learning (Wang et al., 2020) to exploit the knowledge of multiple translation models. However, we argue that these methods either do not fully exploit the knowledge of every individual model (Freitag et al., 2017; Wei et al., 2019) or become computationally expensive during decoding (Wang et al., 2020). Besides, these methods have common limitations that they ignore the usage of word-level knowledge distillation and the complementarity of different decoding strategies.

In this work, we introduce an ensemble learning framework based on knowledge distillation to exploit the knowledge of multiple trained models and aggregate them into a single model. We take these trained models as teachers and the desired model as the student. Unlike the traditional knowledge distillation scenario, the student model has the same size as teachers, and we expect the student to be stronger than every individual teacher after aggregating their knowledge.

[1] We will release the source code and the created SEL training data for reproducibility.

To fit this scenario, we extend the knowledge distillation technique in NMT to the scene of multiple teachers, which gives two methods for ensemble learning, namely Word-level Ensemble Learning (WEL) and Sequence-level Ensemble Learning (SEL). WEL is the extension of word-level knowledge distillation, which uses the predicted probabilities of all teacher models as the target distribution. Similarly, SEL is extended from sequence-level knowledge distillation. Compared to WEL, SEL allows teacher models to have different decoding strategies and hence can combine their strengths. We further introduce Sequence-Word Ensemble Learning (SWEL) to combine the strengths of word-level and sequence-level ensemble learning. In summary, our contributions are:

- We introduce an ensemble learning framework based on knowledge distillation and develop WEL and SEL under this framework.
- We propose to combine the strengths of different decoding strategies in SEL.
- We introduce SWEL to combine sequence-level and word-level ensemble learning.

We conduct experiments on WMT14 English-German translation, WMT16 English-Romanian translation and WMT17 Chinese-English translation. Experimental results show that our approach achieves substantial improvements over the competitive baseline systems and establishes a BLEU score of 31.13 in the WMT14 English-German translation.

2 BACKGROUND

2.1 NEURAL MACHINE TRANSLATION

Deep neural networks with autoregressive encoder-decoder framework have achieved great success on machine translation. Given a source sentence $\boldsymbol{X}$ and a target sentence $\boldsymbol{Y} = \{y_1, ..., y_T\}$, autoregressive NMT models the translation probability sequentially as:

$$p(\boldsymbol{Y}|\boldsymbol{X};\theta) = \prod_{t=1}^{T} p(y_t|y_{<t}, \boldsymbol{X}, \theta), \tag{1}$$

where θ are model parameters, $y_{<t} = \{y_1, \cdots, y_{t-1}\}$ is the translation history and p is the probability distribution. The standard training objective is the cross-entropy loss, which minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{NLL}}(\theta) = -\sum_{t=1}^{T} \log(p(y_t|y_{<t}, \boldsymbol{X}, \theta)). \tag{2}$$

During inference, decoding algorithms such as greedy search and beam search are applied to generate the translation. If there are multiple models available, we can combine these models to get better translations. Assume we have n models $\theta_{1:n}$ and the i-th model predicts the distribution p_i, then the ensemble of n models predicts the distribution e as follows:

$$e(y_t|y_{<t}, \boldsymbol{X}, \theta_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} p_i(y_t|y_{<t}, \boldsymbol{X}, \theta_i). \tag{3}$$

In addition to the mainstream L2R decoding, there are other decoding strategies for NMT that are attracting attention, including R2L decoding, non-autoregressive decoding (Gu et al., 2017), insertion-based decoding (Gu et al., 2019) and synchronous bidirectional decoding (Zhang et al., 2020). Unfortunately, translation probabilities of these models cannot be directly averaged like in Eq.(3), making it difficult to combine their strengths.

2.2 KNOWLEDGE DISTILLATION FOR NMT

Knowledge distillation (Hinton et al., 2015) describes a class of methods for training a smaller student network to perform better by learning from a larger teacher network. Assume that we are learning a classifier $p(y|x;\theta)$ with $|\mathcal{V}|$ classes, and we have access to a learned teacher distribution $q(y|x)$. Instead of minimizing the cross-entropy with the observed data, knowledge distillation uses the teacher distribution $q(y|x)$ as target and minimizes the loss:

$$\mathcal{L}_{\text{KD}}(\theta) = -\sum_{k=1}^{|\mathcal{V}|} q(y=k|x) \times \log p(y=k|x;\theta). \tag{4}$$

In autoregressive L2R neural machine translation, since the standard training objective is the cross-entropy loss, knowledge distillation for multi-class cross-entropy can be applied as follows:

$$\mathcal{L}_{\text{Word-KD}}(\theta) = -\sum_{t=1}^{T}\sum_{k=1}^{|\mathcal{V}|} q(y_t = k|y_{<t}, \boldsymbol{X}) \times \log p(y_t = k|y_{<t}, \boldsymbol{X}, \theta), \tag{5}$$

where $\mathcal{V}$ is the target vocabulary set. The student model is trained to mimic the teacher's prediction at each decoding step, so this method is called Word-level Knowledge Distillation (Word-KD) (Kim & Rush, 2016). Word-KD can be applied to other NMT models like R2L NMT. The limitation is that the student and teacher must have the same decoding strategy. Ideally, we would like the student model to mimic the teacher's actions at the sequence-level:

$$\mathcal{L}_{\text{Seq-KD}}(\theta) = -\sum_{\boldsymbol{Y}} q(\boldsymbol{Y}|\boldsymbol{X}) \times \log p(\boldsymbol{Y}|\boldsymbol{X}, \theta). \tag{6}$$

This method is called Sequence-level Knowledge Distillation (Seq-KD) (Kim & Rush, 2016), which is intractable due to the exponential large search space. The simplest approximation is to replace the teacher distribution q by a one-hot distribution, which has the probability 1 on the translation result of the teacher model. The loss is then:

$$\mathcal{L}_{\text{Seq-KD}}(\theta) \approx -\sum_{\boldsymbol{Y}} 1\{\boldsymbol{Y} = \hat{\boldsymbol{Y}}\} \log p(\boldsymbol{Y}|\boldsymbol{X}, \theta) = -\log p(\hat{\boldsymbol{Y}}|\boldsymbol{X}, \theta), \tag{7}$$

where $\hat{Y}$ is the output from running beam search with the teacher model. In summary, sequence-level knowledge distillation suggests to: (1) train a teacher model, (2) run beam search over the training set with this model and replace the target-side by the translation result, (3) train the student model with cross-entropy on this new dataset.

3 Approach

In this section, we present an ensemble learning framework based on knowledge distillation to aggregate the knowledge of multiple trained models into a single model. We take these trained models as teachers and the desired model as the student. Then we extend the knowledge distillation technique in NMT to the scene of multiple teachers, which gives two methods namely word-level ensemble learning and sequence-level ensemble learning. We then propose sequence-word ensemble learning to combine the strengths of these two methods.

3.1 Word-Level Ensemble Learning

We first introduce Word-level Ensemble Learning (WEL), where the student model aggregates the knowledge of teachers in word-level. Assume we have access to n learned teachers where the i-th teacher has distribution q_i, and the student model has distribution p. Teachers and the student are required to have the same decoding strategy, and we assume them to be autoregressive L2R NMTs. To distill knowledge from multiple teachers, we simply collect the distillation losses from every teacher and use their sum to train the student model:

$$\mathcal{L}_{\text{WEL}}(\theta) = -\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{|\mathcal{V}|} q_i(y_t = k|y_{<t}, \boldsymbol{X}) \log p(y_t = k|y_{<t}, \boldsymbol{X}, \theta). \tag{8}$$

To compute the WEL loss, we can calculate the probability distribution predicted by each teacher and use their average as the target distribution. WEL can be understood as a special case of word-level knowledge distillation, where the teacher model is the ensemble of multiple individual models. Following Kim & Rush (2016), we introduce a hyperparameter α to interpolate the WEL loss and the original cross-entropy loss, which in practice achieves better performance:

$$\mathcal{L}(\theta) = \frac{\alpha}{n}\mathcal{L}_{\text{WEL}}(\theta) + (1-\alpha)\mathcal{L}_{\text{NLL}}(\theta). \tag{9}$$

3.2 SEQUENCE-LEVEL ENSEMBLE LEARNING

In WEL, the constraint is that those teacher models and the student model must have the same decoding strategy, which limits the ability of WEL since the student can only learn from homogeneous teachers. Here, we introduce the Sequence-level Ensemble Learning (SEL), which does not have this limitation so the student model can combine the strengths of different decoding strategies.

Assume that we have access to multiple learned teachers and they can be divided into m groups where each group corresponds to a decoding strategy, and the i-th group contains r_i models. We use q_{ij} to denote the j-th teacher in the i-th group. Similarly, we collect the sequence-level distillation losses from these teachers to train the student model:

$$\mathcal{L}_{\text{SEL}}(\theta) = -\sum_{i=1}^{m}\sum_{j=1}^{r_i}\sum_{\boldsymbol{Y}} q_{ij}(\boldsymbol{Y}|\boldsymbol{X}) \times \log p(\boldsymbol{Y}|\boldsymbol{X},\theta). \tag{10}$$

The loss is intractable due to the exponential large search space. We first propose two methods to approximate the SEL loss, namely single-model approximation and ensemble-model approximation, and then combine them to obtain a better approximation.

Single-Model Approximation The idea of single-model approximation is to approximate the SEL loss by applying the approximation method used in Eq.(7) for every single teacher. We use $\hat{\boldsymbol{Y}}_{\boldsymbol{ij}}$ to denote the beam search result of q_{ij} and use the corresponding one-hot distribution to approximate q_{ij} as follows:

$$\mathcal{L}_{\text{SEL}}(\theta) \approx -\sum_{i=1}^{m}\sum_{j=1}^{r_i}\sum_{\boldsymbol{Y}} 1\{\boldsymbol{Y} = \hat{\boldsymbol{Y}}_{\boldsymbol{ij}}\} \log p(\boldsymbol{Y}|\boldsymbol{X},\theta) = -\sum_{i=1}^{m}\sum_{j=1}^{r_i} \log p(\hat{\boldsymbol{Y}}_{\boldsymbol{ij}}|\boldsymbol{X},\theta). \tag{11}$$

The process of the above single-model approximation can be summarized as: (1) train n teacher models, (2) use every teacher model to run beam search over the training set and concatenate the translation results as target-side, (3) train the student model on this new dataset.

Ensemble-Model Approximation Since teacher models have been divided into groups according to the decoding strategy, models in the same group can be directly aggregated. On this basis, ensemble-model approximation suggests approximating the SEL loss in the group-level. Firstly, we rewrite the SEL loss as follows:

$$\mathcal{L}_{\text{SEL}}(\theta) = -\sum_{i=1}^{m} r_i \sum_{\boldsymbol{Y}} \frac{\sum_{j=1}^{r_i} q_{ij}(\boldsymbol{Y}|\boldsymbol{X})}{r_i} \times \log p(\boldsymbol{Y}|\boldsymbol{X},\theta). \tag{12}$$

Notice that in the above equation, there are m teacher distributions where the i-th teacher distribution is the probability distribution averaged over the i-th group, which motivates us to use the ensemble of the i-th group to approximate this distribution. We use $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ to denote the beam search result of the ensemble of the i-th group and approximate the SEL loss as follows:

$$\mathcal{L}_{\text{SEL}}(\theta) \approx -\sum_{i=1}^{m} r_i \log p(\hat{\boldsymbol{Y}}_{\boldsymbol{i}}|\boldsymbol{X},\theta). \tag{13}$$

The above ensemble-model approximation treats each group as a whole and directly distills knowledge from the group. Models in the i-th group are first aggregated to decode the training set, and then the translation result is repeated for r_i times as target-side. Actually, ensemble distillation (Freitag et al., 2017) is a special case of ensemble-model approximation when all teacher models have the same decoding strategy.

Combination To obtain a better approximation, we combine the above two methods and use both single models and ensembles to approximate the SEL loss. For simplicity, we ignore the coefficient r_i and directly concatenate all translation results to build the training bitext for SEL. In addition, we concatenate the original training bitext and the distillation bitext to train the student model, which shows better performance in practice (Gordon & Duh, 2019; Freitag et al., 2017). In summary, the target side of the final training bitext includes: (1) beam search results of single models, (2) beam search results of ensembles of groups, (3) the original reference.

Algorithm 1 SWEL

Input: training corpus $D = (\mathbf{X}^{1:\mathrm{M}}, \mathbf{Y}^{1:\mathrm{M}})$, teacher models q_{ij}, the number of groups m, the number of SEL teachers n, the number of models r_i
Output: the student model p

initialize the SEL dataset $\tilde{D} = D$
for $i \leftarrow 1$ **to** m **do**
    use the ensemble of group i to decode the training corpus $\mathbf{X}^{1:\mathrm{M}}$, get the result $\hat{\mathbf{Y}}_{\mathbf{i}}^{1:\mathrm{M}}$
    $\tilde{D} = \tilde{D} \cup (\mathbf{X}^{1:\mathrm{M}}, \hat{\mathbf{Y}}_{\mathbf{i}}^{1:\mathrm{M}})$
    for $j \leftarrow 1$ **to** r_i **do**
        use q_{ij} to decode the training corpus $\mathbf{X}^{1:\mathrm{M}}$, get the result $\hat{\mathbf{Y}}_{\mathbf{ij}}^{1:\mathrm{M}}$
        $\tilde{D} = \tilde{D} \cup (\mathbf{X}^{1:\mathrm{M}}, \hat{\mathbf{Y}}_{\mathbf{ij}}^{1:\mathrm{M}})$
set different random seeds to train n models $\{s_1, \cdots, s_n\}$ on the SEL dataset $\tilde{D}$
set $\{s_1, \cdots, s_n\}$ as teacher models and apply WEL to train the student model p on $\tilde{D}$
return the student model p

3.3 Sequence-Word Ensemble Learning

Compared to SEL, WEL can accurately distill knowledge from teachers but cannot aggregate models with different decoding strategies. Besides, SEL can explore the decoding path of teacher models during training, while WEL can only decode in the path of reference sentences and hence suffers from the exposure bias problem (Bengio et al., 2015). This motivates us to combine these two methods so that WEL can also benefit from the larger exploration space and diverse decoding strategies.

We combine these two methods in a pipeline manner and call it Sequence-Word Ensemble Learning (SWEL) since it sequentially applies the sequence-level and word-level ensemble learning. We first obtain the training bitext for SEL and set different random seeds to train n SEL models on the SEL dataset. Then WEL can be directly applied to aggregate these n models into one. The detailed process of SWEL is summarized in Algorithm 1. We denote the number of teachers for SEL as n'. In SWEL, we need to train $n + n'$ teachers in total. If we reverse the pipeline to apply WEL first and SEL later, then we need to train nn' teachers in total, which greatly increases the training cost. Therefore, we use SWEL instead of WSEL to combine these two ensemble learning methods.

4 Related Work

In neural machine translation, the idea of distilling knowledge from multiple models into a single model was first proposed by ensemble distillation (Freitag et al., 2017), which is within the sequence-level knowledge distillation framework but uses the ensemble of multiple models as a teacher. Wei et al. (2019) proposed online distillation to on-the-fly generate a teacher model from checkpoints. Wang et al. (2019); Bi et al. (2019) further proposed to apply multi-agent learning in NMT, where agents learn advanced knowledge from others and work together to improve translation quality. Tan et al. (2019) applied knowledge distillation in multilingual machine translation where the multilingual model distills knowledge from individual models. Nguyen et al. (2019) proposed to diversify the training data by using multiple forward and backward models to augment the original training dataset. Recently, Wang et al. (2020) proposed Transductive Ensemble Learning, which uses all individual models to translate the test set and then fine-tune a strong model on the translated synthetic corpus.

Regarding the combination of different decoding strategies, previous works mainly focus on exploiting R2L decoding to enhance the L2R decoding procedure. Rerank-based approaches (Liu et al., 2016; Sennrich et al., 2016b) utilize R2L translation models to rescore L2R translations. Some researchers (Zhang et al., 2019; Hassan et al., 2018; Yang et al., 2018) attempt to exploit R2L decoding through incorporating a regularization term into the training objective. Zhang et al. (2018) proposed to extract information from hidden states of the R2L decoder to guide the L2R generation. Zhang et al. (2020) proposed synchronous bidirectional inference to generate outputs using both L2R and R2L decoding simultaneously and interactively.

5 EXPERIMENTS

5.1 SETUP

We evaluate our method on three translation tasks of different scales, including WMT16 English-Romanian translation, WMT14 English-German translation and WMT17 Chinese-English translation. For English-German translation and Chinese-English translation, we use both standard tokenized case-sensitive BLEU (Papineni et al., 2002) and detokenized case-insensitive SacreBLEU[2] (Post, 2018) as the automatic metric. For English-Romanian translation, we only report the SacreBLEU since the difference between the two BLEU scores is very small ($<$0.1).

For English-Romanian translation, we use the WMT16 corpus consisting of 600k sentence pairs for the training. We take newsdev2016 and newstest2016 as validation and test sets. For English-German translation, we use the WMT14 corpus consisting of 4.5M sentence pairs for the training. The validation set is newstest2013, and the test set is newstest2014. For Chinese-English translation, the WMT17 corpus consists of 24M sentence pairs, and we remove all duplicates and use the resulted 20.6M sentence pairs for the training. The newsdev2017 is used as the validation set and newstest2017 as the test set. For English-Romanian and English-German translation, we merge the source and target training sets and learn a BPE (Sennrich et al., 2016a) model with 32K merge operations. For Chinese-English translation, we use BPE with 32K merge operations on both sides.

We implemented our approach based on the Transformer model (Vaswani et al., 2017). We use the base version of Transformer for English-Romanian translation and the big version of Transformer for English-German translation. For Chinese-English translation, we conduct experiments on both base and big versions of Transformer. All models are optimized with Adam (Kingma & Ba, 2014). We use 8 NVIDIA V100 GPUs for the training. Following Ott et al. (2018), we accumulate the gradients for 16 batches in English-German translation. The batch size is 4096 for Transformer base and 3584 for Transformer big.

We set different random seeds to obtain multiple teacher models. We report the average BLEU score when we have multiple models with different seeds. For WEL, all teacher models and the student model are autoregressive L2R NMTs. Due to the memory limitation, the number of teacher models is 3 by default. The hyperparameter α to interpolate the two losses is set to be 0.75 for WEL and 0.5 for SWEL. For SEL, the student model is the autoregressive L2R NMT. Teacher models, unless otherwise specified, are 3 L2R NMTs and 3 R2L NMTS. The model size of the student model and teacher models are the same. For SEL and SWEL, we halve the original dropout to match the size of training data.

5.2 MAIN RESULTS

Baseline Results For English-German and English-Romanian translation, we report our system results along with some relative or competitive systems results in Table 1. Comparing with existing systems, we can see that our baseline system, the L2R Transformer, is already very competitive. R2L Transformer is slightly weaker than L2R in the two translation tasks, so it is reasonable to use the L2R model as a student in the following experiments. The ensemble of 6 models brings considerable BLEU improvements, which is accompanied by the cost of large computation requirements.

Ensemble Learning Results Our WEL system distills knowledge from 3 L2R Transformers, which achieves substantial improvements but still underperforms the direct ensemble. SEL additional aggregate the knowledge of 3 R2L Transformers into the L2R student, which even surpasses the performance of ensemble baseline in the WMT14 dataset. Finally, SWEL distills knowledge from 3 SEL models, which combines the strengths of sequence-level and word-level ensemble learning and achieves the best performance. In the WMT14 English-German translation task, SWEL establishes a BLEU score of 31.13, which yields an improvement of 2.23 BLEU over the baseline.

Results on Large-Scale dataset In Table 2, we report our system results on the large-scale Chinese-English translation task. We can see that our methods are still very effective on the large-scale dataset. On both Transformer base and Transformer big, SWEL improves the L2R baseline by nearly 2 BLEU scores.

[2]https://github.com/mjpost/sacreBLEU

	WMT14 En-De		WMT16 En-Ro
	BLEU	SacreBLEU	SacreBLEU
Existing NMT systems			
BIFT (Zhang et al., 2020)	29.21	–	–
Multi-Agent (Bi et al., 2019)	29.67	–	–
Multi-Agent Dual (Wang et al., 2019)	30.05	–	–
ADMIN (Liu et al., 2020)	30.10	29.50	–
MSC (Wei et al., 2020)	30.56	–	–
MultiBranch (Yang et al., 2020)	30.80	29.90	–
Our implementations			
Ensemble Distillation (Freitag et al., 2017)	30.09	29.58	34.31
Data Diversification (Nguyen et al., 2019)	30.31	29.72	34.60
L2R Transformer	28.90	28.30	33.25
Ensemble of 6 L2R Transformers	30.33	29.64	**35.18**
R2L Transformer	28.13	27.61	32.95
Ensemble of 6 R2L Transformers	29.62	29.01	34.98
Our Methods			
WEL (3L2R)	29.78	29.23	34.30
SEL (3L2R+3R2L)	30.80	30.16	34.76
SWEL (3SEL)	**31.13**	**30.46**	35.14

Table 1: BLEU and SacreBLEU scores for English-German and English-Romanian translation on WMT test sets. The number and types of teacher models are indicated in parentheses.

	Transformer base		Transformer big	
	BLEU	SacreBLEU	BLEU	SacreBLEU
L2R Transformer	24.51	24.03	25.35	24.78
Ensemble of 6 L2R Transformers	26.19	25.68	26.84	26.30
R2L Transformer	23.82	23.37	24.28	23.75
Ensemble of 6 R2L Transformers	25.46	24.98	25.49	24.94
WEL (3L2R)	25.29	24.75	25.97	25.41
SEL (3L2R+3R2L)	26.31	25.82	27.00	26.45
SWEL (3SEL)	**26.52**	**25.99**	**27.23**	**26.64**

Table 2: BLEU and SacreBLEU scores for Chinese-English translation on the WMT17 test set. The number and types of teacher models are indicated in parentheses.

5.3 Ablation Study

In sequence-level ensemble learning, we proposed single-model approximation and ensemble-model approximation to approximate the distillation loss, and the original training bitext is also preserved. Here we conduct experiments on the validation set of WMT16 English-Romanian translation to check their influence on SEL by leaving them out respectively. Results are given in Table 3.

System	**SacreBLEU**
L2R Transformer	33.61
SEL (3L2R+3R2L)	35.60
− single-model	35.18
− ensemble-model	35.35
− original bitext	35.23
− 3 L2R	35.22
− 3 R2L	34.93

Table 3: Ablation study for SEL on the WMT16 English-Romanian translation task.

Table 3 shows that all three factors help the model achieve higher performance. Single-model approximation performs better than the ensemble-model approximation, indicating that some useful information may be lost after ensemble. Hence it is necessary for the student model to distill knowledge directly from every single model. The original bitext is also indispensable, which may contain some information ignored by teachers. Removing L2R teachers or R2L teachers will degrade the student performance, indicating that both the two groups of teachers are necessary.

5.4 Number of Teachers

In this section, we study the impact of the number of teachers on student performance on WMT14 English-German translation. For SEL, half of the teacher models are L2R NMTs and the other half are R2L NMTs, and we increased the number of SEL teachers from 0 to 8. For WEL, all teacher models are L2R NMTs. Due to the memory limitation, we only increased the number of WEL teachers from 0 to 4. Figure 1 shows the BLEU scores of WEL and SEL over the number of teachers. We can see that learning from more teacher models generally results in a stronger student model, but the marginal benefits also become smaller. The current number of teachers for SEL and WEL is reasonable, where we can get a strong student model and avoid too much training cost.

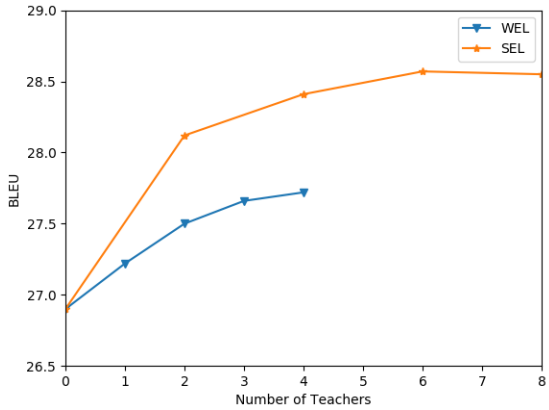

Figure 1: The BLEU scores of WEL and SEL on the validation set of WMT14 English-German translation.

5.5 Complementarity of NMT Models

In SEL, we use 3 L2R NMTs and 3 R2L NMTs as teachers, which is based on the prior knowledge that L2R models and R2L models are complementary, as L2R models tend to generate good prefixes and R2L models tend to generate good suffixes (Zhang et al., 2018; 2019; 2020). However, this conclusion is drawn by observing the translation results and has not been strictly verified. In this section, we aim to find a general method to identify useful teacher models for SEL. We first define a sufficient condition for the complementarity of NMT models and then verify the complementarity among L2R NMTs, R2L NMTs and NATs (Gu et al., 2017). Finally, we explore different variants of SEL and find that complementary teachers can improve the student performance in SEL.

Definition *NMT models A and B are complementary if it satisfies a<b and c>d in the table below.*

	Teacher A	Teacher B
Student A	a	b
Student B	c	d

Table 4: a, b, c and d are BLEU scores of students when applying sequence-level knowledge distillation with different teachers. The target-side includes the teacher translation and original reference.

In the definition above, we give a sufficient condition for the complementarity of two models. Intuitively, when A prefers B as a teacher and B prefers A as a teacher during sequence-level knowledge

distillation, there must be some complementary knowledge in these two models. In the following, we conduct experiments on the WMT14 dataset to see whether L2R Transformer is complementary to the R2L Transformer and NAT. We use the Mask-Predict NAT (Ghazvininejad et al., 2019) as a representative of NAT models.

Table 5: BLEU scores on the validation set of WMT14 English-German translation.

	L2R Teacher	R2L Teacher
L2R Student	27.58	27.90
R2L Student	27.61	26.82

Table 6: BLEU scores on the validation set of WMT14 English-German translation.

	L2R Teacher	NAT Teacher
L2R Student	27.58	26.41
NAT Student	26.53	24.83

From Table 5 and 6, we can see that the L2R Transformer is indeed complementary to R2L Transformer, hence it is reasonable for SEL to learn from 3 L2R teachers and 3 R2L teachers. However, the sufficient condition fails in table 6, suggesting that NAT may be not complementary to the L2R Transformer. We speculate that the main reason is that current NAT models still have a large performance gap with autoregressive models. In appendix A.1, we conduct complementarity experiments on another dataset where the performance gap is smaller, and the conclusion changes.

Teachers	3L2R	6L2R	3L2R+3NAT	3L2R+3R2L	3L2R+3R2L+3NAT
BLEU	27.95	28.09	27.68	28.57	28.34

Table 7: BLEU scores of SEL on the validation set of WMT14 English-German translation.

Table 7 shows the performance of SEL when using L2R Transformers, R2L Transformers and Mask-Predict NATs as teachers. R2L Transformers improve the performance of SEL while NAT models degrade it, which is consistent with the complementarity of models. From the experiments in appendix A.1 and above, we conclude that complementarity can help us find appropriate teachers for SEL since complementary teachers can improve the student performance in SEL.

6 Conclusion

In this paper, we introduce an ensemble learning framework based on knowledge distillation to aggregate the knowledge of multiple teacher models into a single student model. Under this framework, we propose sequence-word ensemble learning, which combines strengths of word-level and sequence-level ensemble learning and achieves substantial improvements over the baseline systems.

References

Umar Asif, Jianbin Tang, and Stefan Harrer. Ensemble knowledge distillation for learning improved and efficient networks. *arXiv preprint arXiv:1909.08097*, 2019.

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

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pp. 1171–1179, 2015.

Tianchi Bi, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. Multi-agent learning for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 856–865, 2019. URL https://www.aclweb.org/anthology/D19-1079.

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

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*, 2017.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. *Proc. Interspeech 2017*, pp. 3697–3701, 2017.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pp. 1607–1616, 2018.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6112–6121, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1633. URL https://www.aclweb.org/anthology/D19-1633.

Mitchell A Gordon and Kevin Duh. Explaining sequence-level knowledge distillation as data-augmentation for neural machine translation. *arXiv preprint arXiv:1912.03334*, 2019.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

Jiatao Gu, Qi Liu, and Kyunghyun Cho. Insertion-based decoding with automatically inferred generation order. *Transactions of the Association for Computational Linguistics*, 7:661–676, 2019.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*, 2018.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. URL https://www.aclweb.org/anthology/D16-1139.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Agreement on target-bidirectional neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 411–416, 2016. URL https://www.aclweb.org/anthology/N16-1046.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. Very deep transformers for neural machine translation. *arXiv preprint arXiv:2008.07772*, 2020.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. Data diversification: An elegant strategy for neural machine translation. *arXiv preprint arXiv:1911.01986*, 2019.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 1–9, 2018. URL https://www.aclweb.org/anthology/W18-6301.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. URL https://www.aclweb.org/anthology/P02-1040.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, 2018. URL https://www.aclweb.org/anthology/W18-6319.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725. Association for Computational Linguistics, 2016a. URL https://www.aclweb.org/anthology/P16-1162.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 371–376. Association for Computational Linguistics, 2016b. doi: 10.18653/v1/W16-2323. URL https://www.aclweb.org/anthology/W16-2323.

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

Xu Tan, Yi Ren, Di He, Qin Tao, Zhao Zhou, and Tie-Yan Liu. Multilingual neural machine translation with knowledge distillation. In *In International Conference on Learning Representations.*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 6000–6010, 2017.

Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, Cheng Xiang Zhai, and Tie Yan Liu. Multi-agent dual learning. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Yiren Wang, Lijun Wu, Yingce Xia, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. Transductive ensemble learning for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Hao-Ran Wei, Shujian Huang, Ran Wang, Xinyu Dai, and Jiajun Chen. Online distilling from checkpoints for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1932–1941, 2019. URL https://www.aclweb.org/anthology/N19-1192.

Xiangpeng Wei, Heng Yu, Yue Hu, Yue Zhang, Rongxiang Weng, and Weihua Luo. Multi-scale collaborative deep models for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 414–426, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.40. URL https://www.aclweb.org/anthology/2020.acl-main.40.

Fan Yang, Shufang Xie, Yingce Xia, Lijun Wu, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. Multi-branch attentive transformer. *arXiv preprint arXiv:2006.10270*, 2020.

Zhen Yang, Laifu Chen, and Minh Le Nguyen. Regularizing forward and backward decoding to improve neural machine translation. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 73–78. IEEE, 2018.

Jiajun Zhang, Long Zhou, Yang Zhao, and Chengqing Zong. Synchronous bidirectional inference for neural sequence generation. *Artificial Intelligence*, pp. 103234, 2020.

Xiangwen Zhang, Jinsong Su, Yue Qin, Yang Liu, Rongrong Ji, and Hongji Wang. Asynchronous bidirectional decoding for neural machine translation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Tong Xu. Regularizing neural machine translation by target-bidirectional agreement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 443–450, 2019.

Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. In *Advances in neural information processing systems*, pp. 7517–7527, 2018.

A APPENDIX

A.1 COMPLEMENTARITY ON ANOTHER DATASET

In section 5.5, we show that L2R Transformer and R2L Transformer are complementary while L2R Transformer and Mask-Predict NAT are not. We speculate that the main reason is that current NAT models have a large performance gap with autoregressive models. In this section, we conduct complementarity analysis on the NIST Chinese-English dataset, which consists of 1.25M sentence pairs for training. We report the standard tokenized case-sensitive BLEU on the NIST02 test set for analysis. We use the base version of Transformer for both L2R and R2L models to make the performance gap smaller. Firstly, we list the BLEU scores of L2R Transformer, R2L Transformer and Mask-Predict NAT in table 8.

Models	L2R	R2L	NAT
BLEU	46.69	45.84	45.73

Table 8: BLEU scores on the validation set of NIST Chinese-English translation.

Table 8 shows that in this dataset, the performance gap between NAT and autoregressive models is not very large. Then we analyze the complementarity between these models by definition. From Table 9 and 10, we find that the L2R Transformer is complementary with both the R2L Transformer and Mask-Predict NAT, which change the conclusion in section 5.5. Therefore, we speculate that NAT models and autoregressive models are indeed complementary, but the complementarity can only be observed when the performance gap is not too large, which raise the necessity of developing better non-autoregressive models.

Table 9: BLEU scores on the validation set of NIST Chinese-English translation.

	L2R Teacher	R2L Teacher
L2R Student	48.40	49.32
R2L Student	48.58	47.72

Table 10: BLEU scores on the validation set of NIST Chinese-English translation.

	L2R Teacher	NAT Teacher
L2R Student	48.40	48.90
NAT Student	47.89	47.16

Teachers	3L2R	6L2R	3L2R+3NAT	3L2R+3R2L	3L2R+3R2L+3NAT
BLEU	48.97	49.03	49.55	50.12	50.44

Table 11: BLEU scores of SEL on the validation set of NIST Chinese-English translation.

Table 11 shows the performance of SEL when using L2R Transformers, R2L Transformers and Mask-Predict NATs as teachers. We can see that both R2L models and NAT models improve the student performance, suggesting that the complementarity can help us to select useful models to serve as SEL teachers.

A.2 PREFIXES AND SUFFIXES

Our approaches achieve substantial BLEU improvements over the baseline systems, but it remains unknown where does our model translate better. In this section, we conduct a study on the generation quality of prefixes and suffixes. We cut the translation and golden reference from the middle and divide them into left and right halves. Then we calculate the left BLEU and right BLEU respectively to roughly measure the generation quality of prefixes and suffixes. We conduct experiments on the validation set of WMT14 English-German and report the results in Table 12.

We can see that WEL improves the translation on both prefixes and suffixes, and the SEL improvements mainly rely on the generation of better suffixes. As L2R models tend to generate good prefixes and R2L models tend to generate good suffixes, it confirms our conclusion that the student model benefit from learning complementary knowledge from teachers.

System	Left BLEU	Right BLEU
R2L Transformer	24.99	23.62
L2R Transformer	25.92	23.47
WEL(3L2R)	26.68	24.19
SEL(3L2R+3R2L)	26.94	25.14
SWEL	27.13	25.40

Table 12: Left BLEU and Right BLEU on the validation set of WMT14 English-German translation.

A.3 SWEL Training Dataset

We proposed to combine SEL and WEL by applying WEL on the SEL dataset. There is an alternative that we train teacher models on the SEL dataset but train the student model on the original dataset. We conduct experiments on the validation set of WMT16 English-Romanian translation to compare these two methods. Figure 2 shows their alpha-BLEU curves.

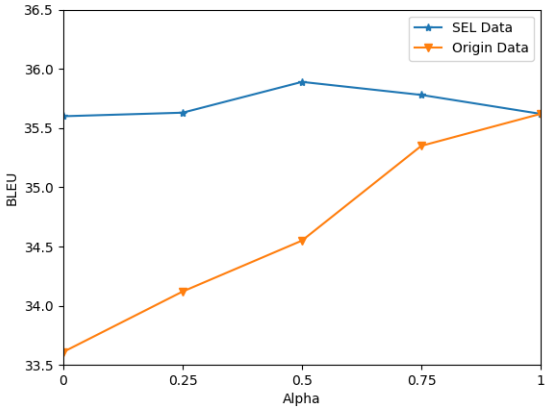

Figure 2: The BLEU scores on the validation set of WMT16 English-Romanian translation.

We can see that when alpha is 1, these two method become equivalent. In other cases, training on the SEL dataset consistently performs better. Besides, it shows stable performance over the hyper-parameter alpha.

A.4 Training Time Analysis

In this section, we conduct a rough analysis on the training time of our proposed methods. We take the training time of the baseline model as the unit of training time.

For WEL, assume that we use 3 teachers, then the pretrain time is 3x. We need these 3 teacher models to do forward calculation to provide their probability distributions, which approximately double the training time. So the total training time is about 3x + 2x = 5x for WEL.

For SEL, assume that we use 3 L2R teachers and 3 R2L teachers, then the training dataset is 9 times the size of the original data. In terms of the training time, it takes about 1.5 times of the original training time to converge, so the total training time is about 6x + 1.5x = 7.5x for SEL.

For SWEL, we need to pretrain 6 teachers first, and then conduct WEL on the SEL dataset. Based on the above discussions, the training time is about 6x + 5 * 1.5x = 13.5x.