

Accurate Knowledge Distillation with n -best Reranking

Hendra Setiawan

Apple

hendra@apple.com

Abstract

We propose utilizing n -best reranking to enhance Sequence-Level Knowledge Distillation (Kim and Rush, 2016) where we extract pseudo-labels for student model’s training data from top n -best hypotheses and leverage a diverse set of models with different inductive biases, objective functions or architectures, including some publicly-available large language models, to pick the highest-quality hypotheses as labels. The effectiveness of our proposal is validated through experiments on the WMT’21 German \leftrightarrow English and Chinese \leftrightarrow English translation tasks. Our results demonstrate that utilizing pseudo-labels generated by our n -best reranker leads to a significantly more accurate student model. In fact, our best student model achieves comparable accuracy to a large translation model from (Tran et al., 2021) with 4.7 billion parameters, while having two orders of magnitude fewer parameters.

1 Introduction

Knowledge Distillation (KD) (Hinton et al., 2015; Buciluă et al., 2006) plays a pivotal role in many machine learning tasks including neural machine translation (NMT). This is evident in recent translation evaluations (Akhbardeh et al., 2021; Kocmi et al., 2022; Agarwal et al., 2023), where the majority of submissions incorporate KD into their training pipelines. Central to this paper is KD’s primary strength, which lies in its ability to utilize a larger *teacher* model to train a smaller *student* model effectively. Consequently, the accuracy of the teacher model correlates with that of the student model.

One well-known yet simple approach for enhancing the accuracy of machine learning models involves model ensembling (Dietterich, 2000), as also has been applied by (Hinton et al., 2015) for KD. In this approach, the underlying models are typically trained on the same datasets but with varying random initializations. Many existing works

applying KD to NMT have adopted this approach, including its dominant variant, namely sequence-level KD (Kim and Rush, 2016), which deploys an ensemble of teacher models to generate the *pseudo-labels* of the student models’ training data. To maintain the simplicity of the inference process when generating the labels, the sequence-level KD often imposes the requirements that the underlying models share the same vocabulary and network architecture. These factors, unfortunately, restrict the types of models that can be included in the ensemble, thereby limiting the avenues for improving the teacher accuracy.

We introduce an n -best reranking approach to extend the vanilla sequence-level KD to allow KD to benefit from a more diverse type of models. N -best reranking is a well-known and powerful approach for boosting translation accuracy, as evident in (Marie et al., 2020; Qian et al., 2021; Tran et al., 2021) just to name a few. Unfortunately, the associated inference cost can be too high for this approach to be deployed in online environments with tight latency constraint. By applying n -best reranking for distillation, the elevated computational cost is shifted to training stage without affecting the latency of the deployed student model, thus sharing a similar motivation as (Yang et al., 2022; Finkelstein et al., 2023).

Our proposed approach involves a two-step process. The first step involves generating a high-quality n -best list from each source sentence. Our initial study indicates a potential gain of almost 10 BLEU on our validation set if we consider hypotheses beyond top-1. The final decision on which hypothesis to be selected is deferred to the second step where we take advantage of models with various architectures, inductive biases, sources of training data and objective functions to rerank the n -best list. To increase model diversity further, we also incorporate open-source large pretrained models in our experiments, tapping into their availability

and increased translation capabilities. This flexibility stands in contrast to the rigid type of teacher models deployed by the vanilla sequence-level KD.

We showcase our reranker’s effectiveness in two scenarios, namely the traditional and iterative KDs. In the former, pseudo-labels are directly utilized for training student models, while the latter, often dubbed as *self-training*, involves an extra step of iteratively retraining the teacher models using the pseudo-labels (Li et al., 2019). In the first scenario, we thoroughly examine the accuracy of student models trained with pseudo-labels generated by our reranker, comparing them to models trained with pseudo-labels from the vanilla sequence-level KD. In the second scenario, we investigate whether our reranker can yield a valuable cascading effect by progressively enhancing teacher models using the pseudo labels from teacher models from the previous iteration. These improved teacher models, in turn, may generate superior pseudo-labels for training more accurate student models.

We also explore efforts that are directed at scaling up our method to distill large-scale training data more efficiently. This includes methods such as *model selection*, where we choose a smaller set of models for distillation that results in only a minimal accuracy drop, and *transfer set reduction*, where we decrease the volume of data to be distilled to only include in-domain samples. More concretely, we conduct extensive experiments on WMT’21 German \leftrightarrow English and Chinese \leftrightarrow English translation tasks. Our final model is as accurate as a large multilingual model with 4.7 billion parameters, despite having only 68 million parameters.

2 Background: Sequence-Level KD

KD trains a student model (p_θ) with the supervision of a more powerful teacher model (q_θ) by minimizing the discrepancy between the prediction of the student model with that of the teacher model. Sequence-level KD, proposed by Kim and Rush (2016), extends KD by minimizing the discrepancy at the level of *sequence* (rather than at token level) by introducing the following loss function:

$$\mathcal{L}_{\text{SEQ-KD}} = - \sum_{t \in \mathcal{T}} q_\theta(t|s) \log p_\theta(t|s)$$

where $t \in \mathcal{T}$ represents the set of all possible sequences that the teacher model can generate from a source sentence (s). Since enumerating all possible sequences is intractable, Kim et al. (2021) approximate the distribution with its mode \hat{t} and arrive at

the following approximation:

$$\mathcal{L}_{\text{SEQ-KD}} \approx -\mathbb{1} \{t = \hat{t}\} \log p_\theta \approx -\log p_\theta(\hat{t}|s),$$

where the mode is obtained via the following inference: $\hat{t} = \arg \max q_\theta(t|s)$. This simple approximation allows NMT to reuse the same standard training pipeline for student model with only a slight modification, namely by substituting the original labels t with pseudo-labels \hat{t} when computing the loss function.

3 N-best Reranking for Distillation

Our n -best reranker formulates $q_\theta(t|s)$ as a weighted log-linear model, which is parameterized by a collection of M models, $\mathcal{M}(s, t) \in \mathbb{R}^M$, and their associated weights $\lambda \in \mathbb{R}^M$. Each model, $\mathcal{M}_i(s, t) \in \mathbb{R}$, assigns a real-valued score that indicates the plausibility of a hypothesis t being the translation of s according to the model, hence we refer to $\mathcal{M}(s, t)$ as scoring models. The score of each scoring model $\mathcal{M}_i(s, t)$ is weighted by λ_i and then combined with all other models to produce the final score. While the models are considered static, their associated weights are *trainable* parameters, learned via discriminative training. We discuss the process in Section 3.1.

To generate pseudo-labels, our reranker applies the following arg max formula:

$$\hat{t} = \arg \max_{t \in \mathcal{N}(s)} \lambda \cdot \log \mathcal{M}(s, t)^\top, \quad (1)$$

where $t \in \mathcal{N}(s)$ refers a hypothesis in an n -best list. We generate the n -best list by running beam search inference with a beam size set to n ; however, it can also be generated using alternative methods such as epsilon sampling. We refer to the models used to generate \mathcal{N} as $\mathcal{G}(s) \subset \mathcal{M}(s, t)$. If $\mathcal{G}(s) = \mathcal{M}(s, t)$ and consists of only one identical translation model, then Eq 1 would revert back to the vanilla sequence-level KD.

3.1 Discriminative Training of λ

To find the optimal λ , we utilize discriminative training and a *tuning set*, which is assumed to be drawn from the same distribution of the test sets. For this paper, we employ the Margin Infused Relaxed Algorithm (MIRA) (Chiang et al., 2008), known for its wide adaptation in Statistical Machine Translation and its ability to handle tens of thousands of inputs. Without loss of generality, we use BLEU (Papineni et al., 2002) to measure translation accuracy in our experiments.

MIRA seeks to find λ that minimizes the following structured hinge loss $\mathcal{L}_{\text{MIRA}}(\lambda)$

$$= \max_{t \in \mathcal{N}} \left[\Delta(t) + \lambda \cdot (\mathcal{M}(s, t)^\top - \mathcal{M}(s, t^*)^\top) \right]$$

where t^* is the *oracle* hypothesis, which refers the hypothesis in the n -best list that attains the highest BLEU score, while $\Delta(t)$ signifies the BLEU differentials of a hypothesis t with the aforementioned oracle hypothesis. Ideally, the optimal λ is achieved when the loss reaches 0, indicating that a clear separation can be established between each non-oracle hypothesis and the oracle hypothesis with a margin proportional to their respective BLEU differentials.

In our experiments, we use a variant of MIRA with an efficient batch-level support, called KB-MIRA (Cherry and Foster, 2012) which can be found in the Moses toolkit (Koehn et al., 2007). It also includes a sparsity-inducing regularization which we utilize for model selection. For a more in-depth discussion on MIRA and its variants, we refer the readers to the cited papers.

3.2 $\mathcal{M}(s, t)$: Models for Scoring n -best List

The efficacy of our n -best reranker depends on the diversity and quality of the deployed scoring models. The log-linear formulation in Equation (1) imposes minimal assumptions about the underlying models, which relaxes the requirements for the models to strictly adhere to probabilistic principles or comprehensively describe the entire translation process. As a result, our reranker is able to accommodate a wide spectrum of models, including heuristics or target-side language models, as long as they assign a relatively meaningful score.

In total, we conduct experiments involving over 50 scoring models for each language pair. For conciseness, we group the models into 8 categories and provide a description for each category, summarized in Table 1. The first four categories include a diverse range of *in-house* translation models, characterized by distinctions in translation directions, generation orders, network architectures, and domain adaptability. It is worth mentioning that most of these models are developed for other exploratory projects and including them as scoring models provides another avenue to reuse them. Meanwhile, the last four categories encompass models that do not strictly pertain to translation models but capture some specific nuances of translation phenomena,

such as the fluency of hypotheses or the level of agreement between hypotheses.

The first category is the *forward translation model* (TM), which includes standard NMT models $p(s|t)$. These models correspond to an autoregressive translation process that generates the translation sequentially one token at a time conditioned on the source sentence and previously generated tokens. The models in this category include NMT models with various well-known architectures, such as Transformer Big (Vaswani et al., 2017), Deep Encoder Shallow Decoder (Kong et al., 2021), Nearest-Neighbor (Khandelwal et al., 2021) and MEGA (Ma et al., 2023).

The second category is the *backward TM*, which includes models that are similar to the forward TM but with different translation direction. In particular, these models focus on modeling the backward translation direction $p(t|s)$, useful to capture how likely the source sentences be the translation of a hypothesis, complementing the forward TM.

The third category is the *right-to-left* TM, which includes models from the first two categories but look at a *reverse* generation order, namely generating tokens in a right-to-left fashion. According to (Liu et al., 2016; Zhou et al., 2019), the left-to-right models are more effective at generating accurate prefixes while the right-to-left models are more effective at generating accurate suffixes.

The fourth category is *domain-adapted* models, which consists of translation models from the previous three categories that we adapt to multiple domains. In our experiments, we simply equate the corpus provenance as the domain. We adopt a tag-based approach and prepend the source sequence with $d \in \{\text{europarl}, \text{commoncrawl}, \text{rapid}, \dots\}$, like in (Johnson et al., 2017; Ha et al., 2017).

The fifth category is the *language model*, consisting of the models that focus on evaluating the fluency aspect of the hypotheses. In our experiments, we train a causal language model with the GPT-2 architecture (Radford et al., 2019) on the target side of our parallel data and the monolingual data. Meanwhile, the sixth category is the *alignment* models, consisting of the models that evaluate the fine-grained correspondences between tokens in the hypothesis and the source sequence. To generate the alignment, we use the IBM model 3 (Brown et al., 1993) from the eflomal toolkit (Östling and Tiedemann, 2016). The seventh category corresponds to the *Minimum Bayes-Risk (MBR) utility function*. Via the models in this category, our

Model category	Formulation	Description
(1) Forward translation model (TM)	$\sum_{i=0}^I \log p(t_i t_{<i}, \mathbf{s})$	Variants: Deep Encoder, MEGA
(2) Backward TM	$\sum_{j=0}^J \log p(s_j s_{<j}, \mathbf{t})$	<i>idem</i>
(3) Right-to-left model	$\sum_{i=0}^I \log p(t_i t_{>i}, \mathbf{s})$	Applied to (1 & 2) as well
(4) Domain-adapted TM	$\sum_{i=0}^I \log p(t_i t_{<i}, \mathbf{s}, d)$	d : domain tag, applied to TM
(5) Monolingual language model	$\sum_{i=0}^N \log p(t_i t_{<i})$	Causal language model
(6) Alignment-based model	$\log P(\mathbf{t} \mathbf{s}, a)$	a : token alignment between y, s
(7) MBR loss function	$\mathcal{U}(\mathbf{t} \mathbf{t}' \in \mathcal{N})$	$\mathcal{U} \in$ BLEU, TER, chrF, etc
(8) Various publicly-available pretrained models, e.g. LASER, mBART, M2M, BLOOM-Z, etc		

Table 1: Categories of models to evaluate hypothesis pair $\mathbf{s}, \mathbf{t} \in \mathcal{N}$ in the n -best list. The first four categories correspond to *in-house* translation models, while the last four correspond to general models.

reranker is able to give preferences to hypotheses that have the higher level of consensus with other hypotheses in the n -best list, as measured by some extrinsic translation metrics. These models infuse our reranker with elements of consensus decoding (Kumar and Byrne, 2004).

Our last category consists of various publicly-available pretrained models. It includes the LASER sentence-embedding model (Artetxe and Schwenk, 2018), the mBART multilingual translation model (Liu et al., 2020), the M2M-100 (Fan et al., 2020) and the NLLB (NLLB Team et al., 2022). It also includes a single dense multilingual model from the WMT21 winning team, namely Facebook AI Research (FAIR) WMT21 (Tran et al., 2021) - currently known as Meta AI Research. Additionally, it includes multilingual large language models from BigScience, namely BLOOMZ and mT0 (Muenighoff et al., 2022). These models are trained with significantly more data and not all of them are explicitly trained to optimize translation objectives. When utilizing these models, we condition them for translation by prepending five translation examples as the prefix (5-shot) like in (Moslem et al., 2023). The sizes of these models vary from 50 million to 10+ billion parameters, which is larger than the models in other categories.

3.3 $\mathcal{G}(\mathbf{s})$: Models for Generating n -best List

The efficacy of our n -best reranking also hinges upon the accuracy and diversity of the n -best list. While an ideal scenario involves deploying all scoring models within $\mathcal{M}(\mathbf{s}, \mathbf{t})$, this proves to be both computationally intensive and impractical, especially considering that not all models explicitly

generate translations, such as language models.

To strike a balance between efficiency and efficacy, we choose to utilize the two specific models, which we call the *L2R* and the *R2L* models. The *L2R* model comprises an ensemble of four Transformer Big models (Vaswani et al., 2017) while the *R2L* model is its right-to-left counterpart. The former belongs to the first category and the latter to the third category described in Table 1. By combining n -best lists from the *L2R* model, specialized at producing accurate prefixes with diverse suffixes, and from the *R2L* model, specialized at generating accurate suffixes with diverse prefixes, we aim to generate highly accurate but diverse n -best lists. Appendix B provides more details about our exploration. For fair comparison, we employ the same *L2R* models as our baseline sequence-level KD experiments.

3.4 Scaling Up n -best Reranking

Deploying the complete set of scoring models to showcase accuracy improvements on a small set of test data is relatively affordable. However, deploying the same complete set to distill the entire training dataset becomes computationally intractable given the scale of the data. Therefore, we describe two efforts to scale up n -best reranking, namely reducing the number of scoring models at distillation time and reducing the number of student model’s training data to distill.

3.4.1 Model Reduction

We seek to find a subset of scoring models, namely $\mathcal{D}(\mathbf{s}, \mathbf{t}) \subset \mathcal{M}(\mathbf{s}, \mathbf{t})$, that provides minimal quality degradation. In our case, we think the goal is attainable since, despite the intended complementarity

of the models, there may be significant redundancy, particularly as the majority of our in-house models are trained on the same data.

Manual selection of $\mathcal{D}(s, t)$ is impractical given the vast number of choices. Instead, we adopt a simple solution by leveraging the discriminatively learned weights λ associated with each scoring model. This approach capitalizes on the regularization term employed by the KB-MIRA optimizer (Section 3.1), offering a convenient and inexpensive way of selecting models that contribute significantly to the task. In our experiments in Section 4.1, we select top 5 models with the highest weights for distillation, reducing the model count in our reranker with minimal accuracy drop.

3.4.2 Transfer Set Reduction

The distillation cost of our n -best reranker is proportional to the size of the so-called *transfer set*, refers to the examples that were distilled and used to train the student model (Hinton et al., 2015). Typically, to maximize accuracy, the transfer set for NMT includes a new set of monolingual data, as suggested by (Edunov et al., 2018). This, in addition to the whole parallel data used also to train the teacher model, significantly increases the distillation cost.

To reduce the distillation cost, thus, we investigate transfer set reduction. Particularly, in our experiments in Section 4.2, we explore using distilled bitext, monolingual data or the combination of both. We found that using only the monolingual data as the transfer set is adequate with no accuracy drop, which leads to a significant saving in distillation time. In addition, we also experiment with a significantly smaller transfer set that consists of the aggregate of multiple in-domain validation sets, used for finetuning a baseline student model, similar to (Finkelstein et al., 2023). Unfortunately, our initial investigation suggests that while it does improve the baseline model’s accuracy, the gain is marginal.

4 Experimental Results

To showcase the efficacy of our proposal, we conduct large-scale experiments on WMT21 German \leftrightarrow English and Chinese \leftrightarrow English translation tasks. Our baseline is the vanilla sequence-level KD (Kim and Rush, 2016) that employs the aforementioned LR models as its teachers. We constrained the student model’s capacity to approximately 68 million parameters, in line with the Transformer *Base* architecture. We use Fairseq

(Ott et al., 2019) for training and inference of our in-house models. More details about these models can be found in Appendix A, including other experimental setup including the bitext used mainly for teacher model training and the monolingual data primarily used for student model training. We use the WMT19 set to learn λ weights for our reranker, the WMT20 set as our validation set and the WMT21 set as our blind test set. For these sets, we use the maximum number of references provided. To report accuracy, we use sacreBLEU (Post, 2018) with the following signature `nrefs:k|case:mixed|eff:no|tok:13a|smooth:exp` where k is the number of reference(s). For our main results, we additionally report chrF (Popović, 2015) with this signature `nrefs:k|case:mixed|eff:yes|nc:6|nw:0|space:no` and COMET22 (Rei et al., 2022) using `wmt22-comet-da` model. For generating n -best list and student model’s hypothesis, we set beam size to 8 and 5 respectively.

In Section 4.1, we first focus on *intrinsic* evaluation, comparing the accuracy of the n -best reranker with that of the sequence-level KD’s teacher models on validation sets. In Section 4.2 and Section 4.3, we then shift to *primary* evaluations where we assess the utility of the pseudo-labels generated by our n -best reranker for training student model and retraining teacher models. We mainly focus on the German \rightarrow English direction and summarize the results for the other language pairs at the end.

4.1 Accuracy of n -best Reranker

Table 2 summarizes the accuracy of our n -best reranker on the German \rightarrow English’s validation set. In the WMT20 set, the baseline system attains a BLEU score of 58.8. This score also represents the score of the top-1 hypothesis in our n -best list since the list is generated by the same model (complemented with its right-to-left counterpart).

In rows *Oracle* and *Anti-Oracle*, we report the accuracy of the best-scoring and worst-scoring hypotheses within our n -best list. Row Oracle shows that the best-scoring hypotheses surpass the top-1 by almost 10 BLEU point, indicating the substantial room for improvement embedded in our n -best reranking approach. Conversely, row Anti-Oracle shows that the gap to the worst-scoring hypotheses is much wider, which is almost 20 BLEU point worse. This underscores the importance of employing robust scoring models, given the risk associated with poor-scoring alternatives.

Description	WMT20
Baseline / Top-1	58.8
Oracle	67.5 †
Anti-Oracle	41.3
n -best Reranker - Full ($ \mathcal{M} = 72$)	60.4 †
n -best Reranker - Select ($ \mathcal{M}^d = 5$)	60.3 †
k NN-MT	59.1
MBR-BLEU	59.3

Table 2: N -best reranker results on WMT20 validation. †implies that the difference is statistically significant with the Baseline at $p < 0.05$.

Using the full set of 72 models (\mathcal{M}), our n -best reranker achieves the BLEU score of 60.4, surpassing the baseline system by 1.6 BLEU point. This outcome underscores the efficacy of our n -best reranker proposal in enhancing model accuracy. We then proceed to apply the model selection strategy described in Section 3.4.1. We pick 5 models with the highest weights, rerun reranking with the same weights (zeroing out the weights of other models) and report the reranker accuracy in the last row. As shown, the accuracy of the n -best reranker with smaller model count is relatively similar to running with the full set of models.

In the last rows of Table 2, we also include two systems from (Yang et al., 2022) and (Finkelstein and Freitag, 2024) for reference. The models from these two system are already included in the scoring models in our n -best reranker experiment. The k NN-MT is based on the vanilla (Khandelwal et al., 2021) trained on the same training data as our baseline. For inference, we set $k = 64$ and $\tau = 100$. For MBR-BLEU, we use our baseline model to generate 260 hypotheses for each source sentences, where we use beam search to generate 4 hypotheses and use epsilon sampling with $\epsilon = 0.02$ to generate the remaining 256 hypotheses, following (Finkelstein and Freitag, 2024). As shown, the accuracy of these two systems are better than the baseline systems, but combining them with other models provide a much better accuracy. Considering the computational cost of k NN-MT and MBR-based in generating hypotheses (discussed in Appendix F), we opt to include these methods as scoring models.

Table 3 compiles the WMT20 accuracy of some models that are eventually utilized in the distillation of the student model’s training data. We rank the models based on the accuracy of each model

when it is used as the *only* model to rerank the n -best list. As shown, the two models utilized for generating the n -best list (\mathcal{G}) are ranked 5 and 13 respectively, but are not selected by our model selection strategy. Interestingly, the models selected for distillation (\mathcal{D}) exhibit considerable variability in terms of ranking, notably excluding the top highest-ranked models. We hypothesize that this is due to redundancy in high-performing models, and the reranker prioritizes model diversity as also suggested by (Gontijo-Lopes et al., 2022). The first model in \mathcal{D} is the single multilingual dense model provided by (Tran et al., 2021), which is the most accurate model. While this model is not their final submission to WMT, it is highly accurate since it is trained on significantly larger training data and consists of 4.7 billion parameters. The remaining four other models in \mathcal{D} come from different model categories, ranging from backward, adapted, R2L and publicly-available models. Note that since the model selection strategy is automatic and non-deterministic, the models chosen for each iteration are dynamic. This is also applicable in other translation pairs.

4.2 N -best Reranking Improves Student

This section investigates the utility of our n -best reranking approach on the downstream task of training student model. We use the reranker with selected models (\mathcal{D}) to generate the pseudo-labels for the whole training data. For our baseline sequence-level KD, we use the L2R model to generate the pseudo-labels. As another baseline, we also include sequence-level Knowledge Interpolation (KI) from (Kim and Rush, 2016), which chooses hypotheses in the n -best list that give the highest BLEU score using the original labels as the references.

As part of scaling up mentioned in 3.4.2, we explore how the accuracy of the student model is impacted by different *transfer sets*. We investigate three configurations, namely *bitext only*, *bitext + monolingual*, and *monolingual only*. Table 4 summarizes the results of our experiments, which contains the accuracy of various student models on WMT21 test set.

In the *bitext only* condition, we only consider the *distilled* parallel data to train the student model. More specifically, we compare the pseudo-labels generated by n -best reranking with three baseline methods: original labels, pseudo-labels obtained through sequence-level knowledge interpolation (KI), and those obtained through sequence-level

Rank	Description	Model Category	\mathcal{G}	\mathcal{D}	WMT20
1	FAIR WMT21 Dense	(8) Public model	-	✓	59.6
5	TransformerBig L2R	(1) Forward TM	✓	-	58.8
13	TransformerBig R2L	(3) Right-to-Left TM	✓	-	58.0
14	TransformerBig $d=cc$	(4) Adapted	-	✓	58.0
19	BigScience mt0-xxl-mt	(8) Public model	-	✓	57.8
32	TransformerBigBwd, R2L $d=rapid$	(2,3,4) Backward, R2L, Adapted	-	✓	54.8
50	TransformerBigBwd, L2R	(2) Backward TM	-	✓	54.7

Table 3: Description of the models used to generate n -best lists (\mathcal{G}) and models selected for distillation (\mathcal{D}), specifically for the first iteration of German \rightarrow English direction, together with their accuracy on WMT20.

Row	Transfer Sets	Baseline	Seq-level KI	Seq-level KD	n -best rerank
1	bitext only (91M)	48.8	49.3	49.6	50.0
2	bitext + mono (155M)	-	-	50.9	52.0
3	mono only (54M)	-	-	50.9	52.2

Table 4: Comparison of BLEU scores on WMT21 test sets between n -best reranking and the three baseline models, including sequence-level knowledge interpolation and distillation, across different configurations of transfer sets.

knowledge distillation (KD). As shown in row 1, the student model trained with the original labels achieved an accuracy of 48.8 BLEU point. Meanwhile, the models trained with pseudo-labels generated through sequence-level KI and KD showed improvements of 0.5 and 0.8 BLEU points respectively, which is in line with previous literature (Kim and Rush, 2016). Our n -best reranker approach leads to even stronger performance, with the student model achieving an accuracy of 50.0 BLEU point. This is a statistically significant improvement of 1.2 BLEU points compared to the baseline.

In the *bitext+mono* condition, we augment the training data for the student model with the distilled monolingual data. Since the monolingual data lack labels, we compare our n -best reranking method only with sequence-level knowledge distillation (KD). The results in row 2 reveal that incorporating the distilled monolingual data significantly improves the accuracy of the sequence-level KD system by approximately 1.3 BLEU points. However, our n -best reranking approach achieves an even greater gain of 2.0 BLEU points, thereby widening the performance gap with sequence-level KD to 1.1 BLEU points. This result highlights the value of incorporating in-domain data as the transfer sets. In our approach, the monolingual data used seems to align with the domain of the evaluation sets, while in contrast, the parallel data are sourced from a broader range of domains.

In row 3, we investigate whether a smaller in-

domain transfer set is more or as effective than a larger mixed-domain one. The results in row 3 reveal marginal gains for both sequence-level KD and our n -best reranking approach when using only the distilled monolingual data as the transfer sets. This result is highly encouraging since we can reduce the distillation time by half without accuracy drop. In any case, our n -best reranking approach leads to a student model that is 3.4 BLEU better than the baseline and 1.3 BLEU better than sequence-level KD. Based on these results, we use monolingual data as the transfer set subsequently.

4.3 Self-Training Teacher Improves Student

Given the substantial accuracy gains obtained by using pseudo-labels generated by n -best reranker in student models, we investigate whether teacher models can derive similar benefit from the use of the same pseudo-labels. Up to now, all the teachers models are trained exclusively from parallel data with original labels that come from a mixed set of domains. In light of this, we conduct a series of experiments retraining the teacher model using pseudo-labels. To manage computational costs effectively, we focus our investigations on retraining the models in \mathcal{G} . This is accomplished through fine-tuning the models, as opposed to retraining them from scratch, and utilizing only monolingual data, excluding the bitext, as the transfer sets. Our rationale for this strategy is detailed in the preliminary experiments, discussed in the Appendix C.

System	$ \theta $	German \rightarrow English			English \rightarrow German		
		BLEU	chrF	COMET22	BLEU	chrF	COMET22
1. Baseline	68M	48.8	67.4	84.8	52.6	68.3	82.0
2. Seq-level KD	68M	50.9 \dagger	68.6	85.7	54.5 \dagger	69.1	83.0
3. n -best (iter 1)	68M	52.2 $\dagger\ddagger$	69.3	85.9	57.4 $\dagger\ddagger$	71.1	84.4
4. n -best (iter 2)	68M	52.7 $\dagger\ddagger$	69.9	85.8	58.3 $\dagger\ddagger$	71.6	84.6
5. n -best (iter 3)	68M	52.8 $\dagger\ddagger$	70.0	86.0	59.1 $\dagger\ddagger$	72.0	84.9
6. FAIR WMT21 Dense	4.7B	52.6 $\dagger\ddagger$	86.3	63.1	59.9 $\dagger\ddagger$	72.7	85.5
7. FAIR WMT21 MoE	52B	53.3 $\dagger\ddagger$	86.5	63.7	62.0 $\dagger\ddagger$	73.5	85.8

System	$ \theta $	Chinese \rightarrow English			English \rightarrow Chinese		
		BLEU	chrF	COMET22	BLEU	chrF	COMET22
1. Baseline	68M	21.2	48.7	60.5	40.3	29.7	79.9
2. Seq-level KD	68M	22.9 \dagger	53.1	73.5	42.5 \dagger	33.1	81.6
3. n -best (iter 1)	68M	28.3 $\dagger\ddagger$	57.4	79.5	43.7 $\dagger\ddagger$	33.3	81.7
4. n -best (iter 2)	68M	29.4 $\dagger\ddagger$	58.1	80.2	45.2 $\dagger\ddagger$	34.9	82.7
5. n -best (iter 3)	68M	30.3 $\dagger\ddagger$	59.4	80.8	45.5 $\dagger\ddagger$	35.2	83.0
6. FAIR WMT21 Dense	4.7B	29.9 $\dagger\ddagger$	60.1	81.9	42.4 $\dagger\ddagger$	34.3	85.2
7. FAIR WMT21 MoE	52B	32.1 $\dagger\ddagger$	60.4	82.2	49.9 $\dagger\ddagger$	39.4	85.2

Table 5: German \leftrightarrow English (top) and Chinese \leftrightarrow English (bottom) results on WMT21 test set, compared to the baseline models and WMT21 models from FAIR. FAIR MoE accuracy is from (Barry Haddow, 2021). Our n -best reranking results are in gray. \dagger implies that the difference with the baseline is statistically significant at $p < 0.05$, while \ddagger implies that the difference with the Seq-level KD is statistically significant at $p < 0.05$.

More specifically, we *fine-tune* the two models in \mathcal{G} for one epoch using the pseudo-labels obtained from the n -best reranker and using monolingual data as the transfer sets. We then retrain the next iteration’s reranker using these models, producing a new set of pseudo-labels for training the student model. It’s worth reiterating that the models selected for distillation \mathcal{D} vary in each iteration. We continue this iterative process *twice* when we typically start observing diminishing gain.

Table 5 provides a summary of our self-training experiments. Focusing on the German \rightarrow English columns, the first three rows of the table are taken from Table 4, reporting the accuracies of the baseline model, the student model trained with sequence-level KD, and the student model trained with pseudo-labels from n -best reranking. The next two rows show the results from our self-training experiments for two iterations. Our experiments show that self-training the teacher models for one iteration can improve the student model accuracy by 0.5 BLEU points (row 4). Our final model, after three iterations, scores 4.0 BLEU points higher than the baseline model and 2.9 BLEU points higher than sequence-level KD. This conclusion is consistent across both chrF and COMET metrics. We also

compare our final model with the winning WMT21 models from FAIR with respect to accuracy and model size, as shown in rows 6 and 7. Performance-wise, our final model is comparable to FAIR’s Dense model, while having fewer parameters. Our model consists of 68 million parameters, while the FAIR model is around 70 times larger.

We also present the experimental results for the English \rightarrow German and the Chinese \leftrightarrow English directions in Table 5. As shown, we observe a gain similar to the one observed in the German \rightarrow English direction where the pseudo-labels from n -best reranker leads to a significantly better student accuracy. These gains remain consistent across multiple metrics, encompassing chrF and COMET22, although given that our reranker is trained to optimize BLEU score, the most pronounced improvement is evident in the BLEU score. Nevertheless, these results affirm our hypothesis that the n -best reranker with robust scoring models can effectively enhance the quality of training data labels.

5 Related Work

Our proposal intersects with many works in various ways. The idea of utilizing n -best reranking to im-

prove accuracy has been extensively investigated as far back as the era of Statistical Machine Translation if not earlier, for example in (Och et al., 2004; Shen et al., 2004; Chiang et al., 2008) and more recently in (Marie et al., 2020; Qian et al., 2021; Tran et al., 2021). In these recent work, n -best reranking incurs significantly higher inference time from running multiple models over the n -best list, thus may not be practical for real-world systems. In contrast, our work makes a practical trade-off by shifting the heavy computational cost of n -best reranking to training data preprocessing without affecting the latency of the deployed model. Our work shares the same motivation as (Yang et al., 2022; Finkelstein et al., 2023), but we consider a larger and more diverse set of models.

The idea of looking at n -best hypotheses for knowledge distillation has been also investigated in the original sequence-level KD paper (Kim et al., 2021), namely sequence-level Knowledge Interpolation which we consider as one of our baseline where the authors propose to approximate the mode with the hypothesis that scores the highest according some translation metrics. However, since this approach requires the ground truth, the application of this variant is limited to distilling parallel data. In contrast, since our n -best reranker is trained on a tune set, our approach is applicable for distilling unlabelled monolingual data.

Our n -best reranker incorporates various models as reranking models. Some of these models have been applied to knowledge distillation. For example, Yang et al. (2022) deploys nearest neighbor machine translation models. Meanwhile, Yee et al. (2019) combines direct models with channel and language models. Currey et al. (2020) trains domain-specific teacher models to distilled in-domain training data for training multi-domain student model. On the other hand, our n -best reranker incorporates significantly larger number of models, including the aforementioned. Also, we deploy these models to score hypotheses, rather than to generate them, which is significantly faster.

Self-training has also been frequently investigated for Machine Translation in statistical and neural era (Li et al., 2019). Recently, it is often dubbed as iterative knowledge distillation and can be found as a winning formula in many evaluation campaigns (Li et al., 2019). In this work, we apply self-training using high-quality pseudo-labels from n -best reranker which produces accurate results.

6 Summary and Future Work

We enhance the sequence-level knowledge distillation (Kim and Rush, 2016) by incorporating n -best reranking. Thus, rather than improving the accuracy of the teacher models by following neural scaling laws alone, our proposed method do so by leveraging a multitude of models with different inductive biases, objective functions or architectures to collaboratively rescore n -best hypotheses and identify the best pseudo-labels. Furthermore, we observed a relatively strong cascading effect, where teacher models finetuned using pseudo-labeled data are more accurate, leading to the generation of more accurate pseudo-labels for the next iteration and resulting in an even more accurate student model. We also explore efficiency efforts to scale up n -best reranking via model selections and transfer set reductions, resulting in a reduction in distillation time. Our final student model demonstrates up to 4.0 BLEU point improvement over baseline systems and is on par with a strong large translation model on German \leftrightarrow English and Chinese \leftrightarrow English translation tasks, despite having only 1/70th the parameters.

For future work, we intend to improve the efficacy of our approach by incorporating more powerful large language models that are finetuned towards translation tasks as well as models that more explicitly capture fine-grained phenomena such as gender or number agreements. To improve the efficiency, we also plan to investigate methods to automatically identify transfer sets at fine-grained sentence level as well as ways to speed up the scoring process further, for instance by utilizing only unnormalized probability score like in (Devlin et al., 2014).

Limitations

While the proposed evaluation framework is language-agnostic, the experiments conducted in this study are limited to two language pairs. Due to its reliance on the availability of models and in-domain monolingual, we cannot guarantee accurate results when applied to language pairs involving a low-resource language pairs. We use numerous pre-trained models with various license terms. While all of them are friendly for non-commercial research purpose, not all of them are not for commercial purpose. Readers should perform their own due diligence.

Ethics Statement

We acknowledge the ethical considerations associated with the n -best reranking approach, which utilizes multiple models to generate pseudo-labels. First of all, we recognize that these models possess their own biases, inherited from the training data, which can potentially perpetuate societal inequalities. Bias in the models can result from biased training data or the inherent limitations of the algorithms used. Despite our best efforts to preprocess and debias the training data, complete elimination of biases is challenging. Second of all, the n -best reranking approach incurs higher computational costs compared to traditional methods. These costs arise from training and maintaining multiple models concurrently. We have implemented mitigation strategies such as model recycling and leveraging publicly available corpora to address these concerns. Furthermore, although we utilize numerous models, it is important to highlight that the majority were not specifically trained for this n -best approach. In fact, many originate from our broader exploratory initiatives, and the n -best reranker serves as a means to repurpose them effectively. Despite of our mitigation efforts, the increased computational burden can limit the accessibility and affordability of the approach, particularly for researchers or organizations with limited resources.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. **FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN**. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. **Findings of the 2021 conference on machine translation (WMT21)**. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2018. **Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond**. *CoRR*, abs/1812.10464.

Barry Haddow. 2021. WMT21 News Systems and Evaluations. <https://github.com/wmt-conference/wmt21-news-systems/blob/main/scores/automatic-scores.tsv> (May 5, 2023).

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. **Massive exploration of neural machine translation architectures**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1451, Copenhagen, Denmark. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. **The mathematics of statistical machine translation: Parameter estimation**. *Computational Linguistics*, 19(2):263–311.

Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. **Model compression**. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, page 535–541, New York, NY, USA. Association for Computing Machinery.

Colin Cherry and George Foster. 2012. **Batch tuning strategies for statistical machine translation**. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada. Association for Computational Linguistics.

David Chiang, Yuval Marton, and Philip Resnik. 2008. **Online large-margin training of syntactic and structural translation features**. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 224–233, Honolulu, Hawaii. Association for Computational Linguistics.

- Anna Currey, Prashant Mathur, and Georgiana Dinu. 2020. [Distilling multiple domains for neural machine translation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4500–4511, Online. Association for Computational Linguistics.
- Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. [Fast and robust neural network joint models for statistical machine translation](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland. Association for Computational Linguistics.
- Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems, MCS '00*, page 1–15, Berlin, Heidelberg. Springer-Verlag.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. [Understanding back-translation at scale](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Man-deep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. [Beyond english-centric multilingual machine translation](#). *arXiv preprint*.
- Mara Finkelstein and Markus Freitag. 2024. [MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods](#). In *The Twelfth International Conference on Learning Representations*.
- Mara Finkelstein, Subhajit Naskar, Mehdi Mirzazadeh, Apurva Shah, and Markus Freitag. 2023. [Mbr and qe finetuning: Training-time distillation of the best and most expensive decoding methods](#).
- Raphael Gontijo-Lopes, Yann Dauphin, and Ekin Dogus Cubuk. 2022. [No one representation to rule them all: Overlapping features of training methods](#). In *International Conference on Learning Representations*.
- Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2017. [Effective strategies in zero-shot neural machine translation](#). In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 105–112, Tokyo, Japan. International Workshop on Spoken Language Translation.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#).
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. [Nearest neighbor machine translation](#). In *International Conference on Learning Representations (ICLR)*.
- Beomsu Kim, Seokjun Seo, Seungju Han, Enkhbayar Erdene, and Buru Chang. 2021. [Distilling the knowledge of large-scale generative models into retrieval models for efficient open-domain conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3357–3373, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yoon Kim and Alexander M. Rush. 2016. [Sequence-level knowledge distillation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.
- Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. [Findings of the 2022 conference on machine translation \(WMT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. [Moses: Open source toolkit for statistical machine translation](#). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07*, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Xiang Kong, Adithya Renduchintala, James Cross, Yuqing Tang, Jiatao Gu, and Xian Li. 2021. [Multilingual neural machine translation with deep encoder and multiple shallow decoders](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1613–1624, Online. Association for Computational Linguistics.
- Shankar Kumar and William Byrne. 2004. [Minimum Bayes-risk decoding for statistical machine translation](#). In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*:

- HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao, and Jingbo Zhu. 2019. [The NiuTrans machine translation systems for WMT19](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266, Florence, Italy. Association for Computational Linguistics.
- Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. [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*, pages 411–416, San Diego, California. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#).
- Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2023. [Mega: Moving average equipped gated attention](#). In *The Eleventh International Conference on Learning Representations*.
- Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. [Combination of neural machine translation systems at WMT20](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 230–238, Online. Association for Computational Linguistics.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. [Adaptive machine translation with large language models](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. [Crosslingual generalization through multitask finetuning](#).
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Meja-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. [No language left behind: Scaling human-centered machine translation](#).
- Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. [A smorgasbord of features for statistical machine translation](#). In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 161–168, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Robert Östling and Jörg Tiedemann. 2016. [Efficient word alignment with Markov Chain Monte Carlo](#). *Prague Bulletin of Mathematical Linguistics*, 106:125–146.
- 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*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Lihua Qian, Yi Zhou, Zaixiang Zheng, Yaoming Zhu, Zehui Lin, Jiangtao Feng, Shanbo Cheng, Lei Li, Mingxuan Wang, and Hao Zhou. 2021. [The voltrans GLAT system: Non-autoregressive translation meets WMT21](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 187–196, Online. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins.

2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. [Discriminative reranking for machine translation](#). In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. [Facebook AI’s WMT21 news translation task submission](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.
- 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*, volume 30. Curran Associates, Inc.
- Zhixian Yang, Renliang Sun, and Xiaojun Wan. 2022. [Nearest neighbor knowledge distillation for neural machine translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5546–5556, Seattle, United States. Association for Computational Linguistics.
- Kyra Yee, Yann Dauphin, and Michael Auli. 2019. [Simple and effective noisy channel modeling 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)*, pages 5696–5701, Hong Kong, China. Association for Computational Linguistics.
- Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019. [Synchronous bidirectional neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 7:91–105.

A Experimental Setups

We follow the experimental setup of the WMT21 news translation task, particularly the constrained track to train our in-house models. For German \leftrightarrow English directions, our parallel data are composed of Europarl v10, ParaCrawl v7.1, Common Crawl, News Commentary v16, Wiki Titles v3, Tilde Rapid and WikiMatrix. For Chinese \leftrightarrow English directions, our parallel data are composed of Paracrawl v7.1, News Commentary v16, Wiki Titles v3, UN Parallel Corpus v1.0, CCMT and WikiMatrix. For monolingual data, we use the 2021 subsets of News Crawl. We deduplicate and preprocess the data using the M2M-100 (Fan et al., 2021) processing scripts¹. For training our in-house teacher models, we run up to 80 thousand updates, while for training the student model, we run up to 30 thousand updates. For finetuning teacher models, we run one epoch of updates.

Table 6 summarizes the data sizes for different splits of the two language pairs.

	Bitext	Mono	Valid	Test
De \rightarrow En	91M	38M	785	1000
En \rightarrow De	91M	37M	1418	1002
Zh \rightarrow En	54M	32M	2000	1948
En \rightarrow Zh	54M	37M	1418	1002

Table 6: Sizes of Splits used in the paper, where Valid refers to WMT20 and Test refers to WMT21.

B Pilot Study for Models for n -best Generation

Figure 1 shows the BLEU scores for the top-1 and oracle hypotheses of n -best list with different N from 1 to 32 on our tune set. As shown, the BLEU score for the top-1 hypotheses marginally improves when we increase the beam size from 1 to 4 but then it plateaus, which is consistent with Britz et al. (2017)’s finding. This suggests that increasing the beam size may not benefit the original sequence-level KD. In contrast, the oracle BLEU score improves monotonically with larger beam size, where the gap for $N > 8$ is more than 10 BLEU points and growing. This gap speaks to the potential for our proposed n -best reranking. Compared to L2R, the n -best list’s oracle score from L2R+R2L is around 2-3 BLEU points higher. We equate \mathcal{G} to

¹https://github.com/facebookresearch/fairseq/tree/main/examples/m2m_100

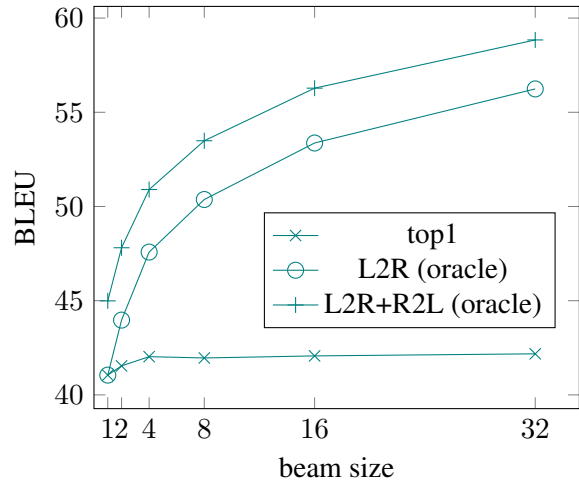


Figure 1: BLEU scores for top-1 and oracle hypotheses of WMT19 with different beam size

L2R+R2L with beam size of 8 since its accuracy is better than doubling the beam size of L2R setup with additional parallelization benefits.

C Pseudo-Labels from N -best reranking for Self-Training

We conduct a pilot study on *one* of the L2R models, which is part of \mathcal{G} , to inform our decisions on two aspects: 1) determining which transfer sets to utilize, and 2) determining whether it is necessary to retrain the teacher model from the scratch or if finetuning proves to be sufficient. The results of this pilot study are summarized in Table 7. For finetuning, we only run one epoch, while for retraining we run around 50 epochs (up to 80 thousands update). The baseline accuracy of training this teacher model using the original bitext is 57.4 point, as indicated in the first row of column *Baseline*. The *Retrain* column shows that training the teacher model with the same bitext, but with pseudo-labels, resulted in a gain of 0.6 BLEU point. As shown in the subsequent rows (bitext+mono and mono only), adding the distilled monolingual data to the transfer sets or using them alone result in a stronger gain of around 1.5 BLEU points, which is consistent with our finding in the student model training.

Comparing the Retrain and Finetune columns, we observe that the accuracy of finetuned models is on par with the model trained from scratch. These results are encouraging because we can obtain a teacher model that is 2.1 BLEU points more accurate with minimal training FLOPs via finetuning and using the smallest transfer set. We conduct similar experiments using pseudo-labels from

Transfer Sets	Baseline	Retrain	Finetune
bitext only	57.4	58.0	58.0
bitext+mono	-	59.5	59.2
mono only	-	59.5	59.5

Table 7: WMT20 scores of a teacher model trained with pseudo-labels from n -best reranking with different transfer sets (rows) and training regime (columns).

sequence-level KD and discuss it in Appendix D. Although a similar trend is observed, the resulting gain from sequence-level KD is smaller.

D Pseudo-Labels from Sequence-level KD for Self-Training

We report the results for self-training teacher model using the pseudo-labels from sequence-level KD in Table 8. As shown in row *bitext only*, retraining teacher models with these pseudo-labels leads to a degradation. Including the monolingual data as the transfer sets helps to improve the accuracy as shown in row *bitext+mono* and *mono only*. Comparing columns Retrain and Finetune, we observe that finetuning can achieve a similar accuracy gain as the full retraining, which is similar to what we observe in finetuning experiments using pseudo-labels from n -best reranking. Comparing with using pseudo-labels from n -best reranking reported in Table 7, self-training using pseudo-labels from sequence-level KD gives smaller accuracy gain than self-training using pseudo-labels from n -best reranking.

	Baseline	Retrain	Finetune
bitext only	57.4	57.3	57.1
bitext+mono	-	58.3	57.8
mono only	-	58.3	58.1

Table 8: WMT20 BLEU scores of a teacher model trained with pseudo-labels from *sequence-level KD* with different transfer sets (rows) and training regime (columns).

E Effects of Pseudo-Labels on Different Model Architectures

This section describes the efficacy of pseudo-labels generated by n -best reranker on the teacher models, beyond the student model described in the main paper. Table 9 details the accuracy of teacher models

with and without reranking along with the accuracy of the corresponding student models, focusing on the German \rightarrow English WMT21 test set. For the student models, we copy the numbers from the German \rightarrow English section of Table 5. At iteration 0, the student model is trained with the original labels of the bitexts, while at the later iterations, the student is trained with monolingual data with pseudo-labels generated using n -best reranker at the corresponding iteration. For column Top-1, we report the accuracy of the generating models, which refer to an ensemble of 4 models with Transformer Big architecture, each consisting of around 310 million parameters. At iteration 1, these teacher models are trained with original labels of the bitexts, and at later iterations, they are trained with monolingual data with pseudo labels generated by the reranker at the previous iteration.

As shown, there is a 2 BLEU point gap between the student model (row iter 1; column Student) and the teacher models (row iter 1; column Top-1) when the two models are trained with parallel data with original labels. As shown, the n -best reranker improves the teacher accuracy by +2.5 BLEU point (from 50.8 to 53.3). When this reranker is used to generate the pseudo-labels of the monolingual data for student model training, the student model’s accuracy increases up to 52.2 BLEU score (row iter 2; column student). When the same pseudo-labels are used to train the teacher models, the accuracy of teacher model’s next iteration increases up to 52.5 BLEU point (row iter2 and column Top-1). A similar trend but with less significant improvement is also observed for iteration 2 and iteration 3. This result demonstrates that the accuracy gain observed in the student model is also observed in the teacher models, which is an order magnitude larger. Additionally, it also shows that the accuracy gap between teacher and student models is smaller with the combination of self-training and n -best reranking.

F Distillation Cost

Table 10 presents the distillation cost incurred for distilling a sample of 10,000 German sentences utilizing the n -best reranker outlined in Table 1. Since our n -best reranking consists of many parallelizable components, we detail the costs in terms of *parallel* and *serial* hours. Parallel hours depict a scenario in which all computing resources are accessible simultaneously, while serial hours de-

Iter	Student	Teacher		
		Top-1	Reranked	Δ
1	48.8	50.8	53.3	+2.5
2	52.2	52.5	53.6	+1.1
3	52.7	53.0	54.0	+1.0
4	52.8			

Table 9: WMT21 BLEU scores of teacher and student models across different iteration for the WMT21 German-English test set.

Step	Parallel Hours	Serial Hours
Generating n -best	00:48	01:33
TransformerBig L2R	00:48	
TransformerBig R2L	00:45	
Scoring	02:08	06:22
FAIR WMT21 Dense	01:55	
TransformerBig $d=cc$	00:43	
BigScience mt0-xxl-mt	02:08	
TransformerBigBwd,R2L, $d=rapid$	00:51	
TransformerBigBwd,L2R	00:45	
arg max	00:10*	
n -best reranking total	03:06	08:05
k NN-NMT	14:54	
MBR-BLEU	15:28	

Table 10: GPU hour breakdown of a German-English distillation process for a sample of 10,000 German sentences using n -best reranker described in Table 1. Last two rows report the distillation cost for two other methods. * refers to CPU hours

pick a scenario where only one resource is available at a time. The actual wallclock time is contingent upon the condition of the compute cluster condition, which impacts the actual level of parallelism. Apart from the reranking step which computes the final cost of each hypothesis, all the cost in Table 10 refers to GPU hours.

As shown, generating the n -best list takes around one and a half hour to generate using the two generation models. For the baseline knowledge-distillation, utilizing solely the L2R model, the pseudo-labels can be generated in less than an hour. The majority of cost for n -best stems from the scoring step, involving 5 models. The cost for each model correlates with the model size. Scoring using a 13 billion models (BigScience mt0-xxl-mt)

takes around 2 hours while scoring using an ensemble of 4 models with 655 million parameters (TransformerBig) takes around 45 minutes. The parallel cost for the scoring step is around 2 hours, which represents an optimal situation where the computing resources are available to score using all models, which is equal to the time needed for the slowest model. Meanwhile, the serial cost for this step takes up eight hours, which represents a less than ideal situation where each model must wait for resources sequentially. The last step (arg max) is to takes the scored n -best and rerank it according to the learned weights. This step only requires CPU and the cost is negligible compared to other steps. In practice, the n -best reranking takes between 3:06 to 08:05, incurring around 4.0x to 10x times distillation cost than the baseline sequence-level KD.

In the last two rows, we report the distillation costs from two related work, namely k Nearest Neighbor- $(k$ -NN NMT) from (Yang et al., 2022) and Minimum Bayes Risk decoding (MBR-BLEU) from (Finkelstein and Freitag, 2024). For k NN-NMT, we generate the pseudo-labels with beam size of 8 similar with the baseline and set k to 64 neighbors and the temperature $\tau=100$. For MBR-BLEU, following (Finkelstein and Freitag, 2024), we generate 260 hypotheses for each source sentences by generating 4 hypotheses via beam search with beam size of 4 and the remaining 256 hypotheses via epsilon sampling with $\epsilon = 0.02$. As shown, the distillation costs for these two related methods are substantially larger than our n -best reranker approach. The cost associated with k NN-NMT is consistent with the conclusion of (Khandelwal et al., 2021) where they reported two order magnitude slower inference speed than their baseline. Meanwhile, the cost associated with MBR-BLEU is due to the high number hypotheses generated in the second decoding stage, which requires us to reduce the effective batch size significantly during inference time.

Lastly, while distilling with n -best reranking introduces a notable increase in computational cost, it however offers a substantial improvement in accuracy. Thus, it is essential to acknowledge that while the cost of n -best reranking may be higher, it pales in comparison to the labor-intensive process of manually creating new parallel data. Therefore, the expense incurred by n -best reranking should be considered within the context of its significant accuracy gains and the resource-intensive alternative of generating new parallel data manually.