Norm-based Noisy Corpora Filtering and Refurbishing in Neural Machine Translation

Yu Lu[1,2] **and Jiajun Zhang**[1,2]*
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{yu.lu, jjzhang}@nlpr.ia.ac.cn

Abstract

Recent advances in neural machine translation depend on massive parallel corpora, which are collected from any open source without much guarantee of quality. It stresses the need for noisy corpora filtering, but existing methods are insufficient to solve this issue. They spend much time ensembling multiple scorers trained on clean bitexts, unavailable for low-resource languages in practice. In this paper, we propose a norm-based noisy corpora filtering and refurbishing method with no external data and costly scorers. The noisy and clean samples are separated based on how much information from the source and target sides the model requires to fit the given translation. For the unparallel sentence, the target-side history translation is much more important than the source context, contrary to the parallel ones. The amount of these two information flows can be measured by norms of source-/target-side context vectors. Moreover, we propose to reuse the discovered noisy data by generating pseudo labels via online knowledge distillation. Extensive experiments show that our proposed filtering method performs comparably with state-of-the-art noisy corpora filtering techniques but is more efficient and easier to operate. Noisy sample refurbishing further enhances the performance by making the most of the given data[1].

1 Introduction

Neural machine translation (NMT) has achieved significant progress with help from large parallel corpora for training (Tiedemann, 2012; Smith et al., 2013). These data are typically extracted from the web without much control over the quality, which presents misalignment, wrong languages, too many numbers or URLs, etc. In this case, noisy corpora filtering holds a critical research area to prevent noisy bitexts from degrading the generalization performance of NMT (Khayrallah and Koehn, 2018).

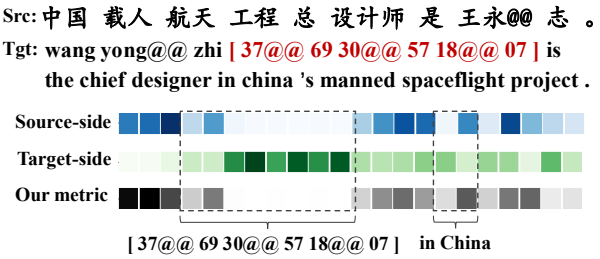

Figure 1: An example of the required source-/target-side information when the model fits an unparallel Zh⇒En sentence pair (words in red are real noisy segments). The amount of information on each side is counted by norms of corresponding context vectors, positively correlated to the darkness of color blocks.

Much effort has been devoted to this field with the promotion of a WMT shared task for parallel corpus filtering. However, prior work is difficult to operate in practice due to two drawbacks. (1) *High time and computational cost*. Their good performance relies on the ensembling of multiple large-scale scorers, which involves costly pre-training and fine-tuning (Esplà-Gomis et al., 2020; Lu et al., 2020). (2) *Dependence on clean bitexts*. The training of above scorers needs clean parallel sentence pairs as positive samples, which are scarce for low-resource languages in real-world applications.

This paper introduces a norm-based noisy sample filtering and refurbishing method, which avoids extra clean bitexts and heavy scorers. We distinguish unparallel sentence pairs from others based on observed model behaviors during the training of NMT. Generally, the model captures two aspects of information to predict the given translation, the source-side context information from the encoder and the target-side one from history translations in the decoder. For unparallel sentence pairs, provided translations are partially or entirely unrelated to the source sentence. In this case, the NMT model behaves as a language model, which requires excessive target-side information to fit noises. Thus,

*Corresponding author.
[1]https://github.com/yulu-dada/Norm_NoisyFiltering

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5414 - 5425
December 7-11, 2022 ©2022 Association for Computational Linguistics

we use the information ratio of the source to the target side as the criterion to filter noises.

Specifically, we observe that a greater vector norm implies richer context information captured by the model. Thus, we calculate the amount of each information flow by norms of corresponding context vectors. This metric is easy to obtain in the training process and sufficient to model how much information is encoded on each side. We take Figure 1 as an example. When the model predicts the content word "china", the norm of the source-side context information is more significant than that of the target side. The opposite situation is in generating the function word "in". However, the quantity of target-side information appears to be exceptionally great for the noisy fraction, which presents a deeper color than others, leading to a lower score than correct translations under our estimation.

We further propose to refurbish discovered noisy samples by generating pseudo labels via online knowledge distillation. By doing this, the source sentence in Figure 1 is regarded as the monolingual data to complement limited clean bitexts. Throughout the whole, we incorporate filtering and refurbishing into the training of NMT rather than separating data filtering and training, thus considerably improving computational efficiency.

We validate the effectiveness of our approaches on Transformer-based NMT (Vaswani et al., 2017), including the WMT2020 shared task for parallel corpus filtering (Km⇒En and Ps⇒En) and our in-house web-crawled datasets (He, Id, Pt, Ko, and Es⇒Zh). Empirical results show that our proposed method performs comparably with SOTA noisy corpora filtering approaches. Refurbishing noisy samples further substantially boosts the performance. Detailed analyses show that our metric can reflect the alignment extent at word and sentence levels.

The contributions of this paper are three-fold:

- We propose to use norms of source- and target-side context vectors to represent the amount of information flowing from each side. We find that the model needs excessive target-side information to fit the unparallel sentence pair, which is the basis of the following work.
- We propose a norm-based noisy corpora filtering method by calculating the information ratio from the source to the target side. It is experimentally efficient and effective under the condition of no extra data and costly scorers.

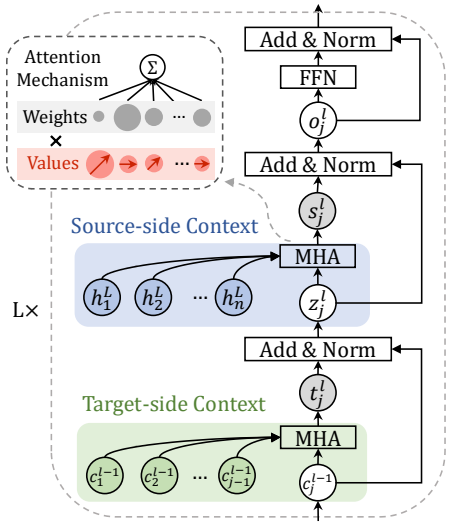

Figure 2: Structure of l-th decoder block in Transformer. The model first sees history translations and then source-side contexts by attention mechanism and obtains two context vectors t_j^l and s_j^l. In the enlarged view of MHA, the arrows in the circles represent the corresponding vectors. The sizes of circles illustrate the values of attention weights or the vectors' norms.

- We propose to refurbish discovered noisy samples by producing pseudo labels via online knowledge distillation, which makes the most of the corpora and further boosts the performance.

2 Background

In this section, we first briefly introduce a mainstream NMT framework, Transformer, with a focus on how to capture source- and target-side contexts. We then present how the vector norm serves as an indicator of diverse features, which motivates us to count how much information is encoded in context vectors of each side based on vector norms.

2.1 Transformer-based NMT

The Transformer is an encoder-decoder framework which alternately looks over source- and target-side contexts to make prediction. The encoder with L layers transforms an input $\mathrm{x} = \{x_1, x_2, ...x_n\}$ to a sequence of hidden states $\boldsymbol{h}^L = \{h_1^L, h_2^L, ...h_n^L\}$, from which the decoder predicts the probability of a target sentence $\mathrm{y} = \{y_1, y_2, ...y_m\}$.

$$P(\mathrm{y}|\mathrm{x}) = \prod_{j=1}^{m} p(y_j|\mathrm{y}_{<j}, \mathrm{x}) = \prod_{j=1}^{m} p(y_j|c_j^L) \quad (1)$$

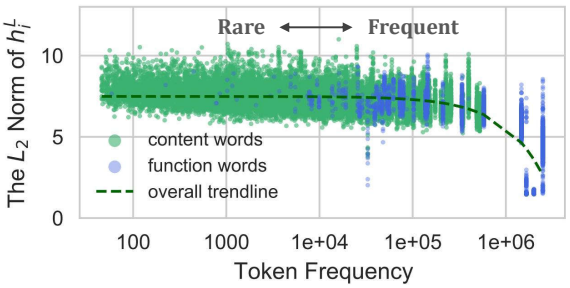

Figure 3: The L_2 norm of h_i^L versus token frequency of all Chinese words in the LDC Zh⇒En vocabulary labeled content words (green dots) and function words (blue dots). This produces a downward trendline.

where $y_{<j}$ is a partial translation. c_j^L denotes the j-th hidden state in the L-th decoder layer. We take the l-th decoder layer as an example in Figure 2. The model first attends to the history translation $\boldsymbol{c}_{<j}^{l-1}$ by multi-head attention (MHA) and obtains the target context vector t_j^l.

$$t_j^l = \mathrm{MHA}(\boldsymbol{c}_{<j}^{l-1}, c_j^{l-1}) \tag{2}$$

where MHA enables dynamically selecting relevant tokens by assigning different attention weights. t_j^l is then transformed to z_j^l by layer normalization (Ba et al., 2016) and residual network (He et al., 2016). The model later looks at the source-side context to obtain the source context vector s_j^l:

$$s_j^l = \mathrm{MHA}(\boldsymbol{h}^L, z_j^l) \tag{3}$$

Two-side information is mixed up to calculate the next-layer hidden state c_j^l.

$$\begin{aligned} o_j^l &= \mathrm{LayerNorm}(s_j^l + z_j^l) \\ c_j^l &= \mathrm{LayerNorm}(o_j^l + \mathrm{FFN}(o_j^l)) \end{aligned} \tag{4}$$

where FFN denotes a feedforward neural network.

2.2 Norm-based Word Importance Measurement

As the key element in the NMT model, word representations capture rich semantic features. Schakel and Wilson (2015) report that the L_2 norm of word vectors learned in the word2vec model (Mikolov et al., 2013) is informative, where words with low frequency or diverse contexts are more likely to be assigned higher norms. Liu et al. (2020) state that the norm of word embeddings in the NMT model is also a good proxy of word importance.

Here, we extend to hidden states attended by the attention mechanism ($\boldsymbol{h}^L$ produced by the encoder and c_j^{L-1} in the decoder). As shown in Figure 3, norms of $\boldsymbol{h}^L$ shift downwards with the frequency increasing. Specifically, norms of content words are relatively higher than function words. It suggests that the rare and informative words obtain a high norm of $\boldsymbol{h}^L$, which stays consistent with results of c_j^{L-1} as given in Appendix A. Thus, norms of hidden states that the attention mechanism looks at can indicate word importance.

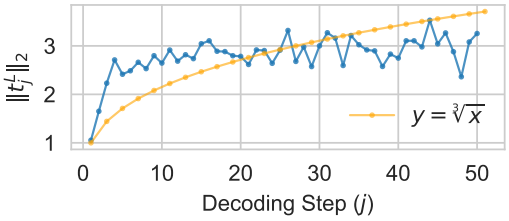

Figure 4: The L_2 norm of target-side context vectors t_j^L at L-th layer with respect to the decoding step j. We use the function $y = \sqrt[3]{x}$ to fit the changing trend.

As context vectors are formed as a weighted sum of those hidden states via attention mechanism, the derived context vectors' norms would grow in line with the model's increasing attention towards informative words with higher norms. These observations motivate us to evaluate how much information is encoded in context vectors from this point.

3 Methodology

We aim to detect and refurbish noisy sentence pairs by observing how the model predicts each token. A sentence pair is potentially misaligned if the model depends heavily on history predictions rather than the source sentence to fit given translations. To this end, we first introduce a norm-based measurement to count the amount of information extracted from the source and target side (Section 3.1). Then, we show how to use this metric to filter noisy samples (Section 3.2), which are further refurbished by producing pseudo labels via online knowledge distillation (Section 3.3).

3.1 Norm-based Source- and Target-side Information Measurement

As shown in Figure 2, the model repeatedly collects information from the source sentence ($\boldsymbol{h}^L$) and history translation ($\boldsymbol{c}_{<j}^{l-1}$) to calculate context vectors by attention mechanisms. Specifically, it computes a weighted sum of hidden states, the norm of which indicates word importance as stated in section 2.2. If the model pays more attention to content words

with greater norms, the norm of obtained context vector would correspondingly increase, and vice versa. Thus, we can use the norm of the source and target context vector ($\|s_j^l\|_2$ and $\|t_j^l\|_2$) to count the amount of information extracted from two sides.

However, directly comparing $\|t_j^l\|_2$ at different steps is unfair, for more history translations are available for the model in the later steps. In this case, the decoder has access to more content words and gets a high-norm context vector. As shown in Figure 4, $\|t_j^L\|_2$ rapidly increases at first, and then the growth slows down. The overall trend is similar to the function $y = \sqrt[3]{x}$. Thus, we normalize the norm of the target context vector with $\sqrt[3]{j}$. Here, we extract context vectors from the L-th layer and design the metric as:

$$\gamma_j = \frac{\|s_j^L\|_2}{\|t_j^L\|_2 / \sqrt[3]{j}} \tag{5}$$

which is positively related to how much the source sentence is relied on to make predictions. Different values of γ_j indicate different cases:

- If γ_j is big, the model mainly depends on the source sentence x to predict y_j, which may be nouns, verbs, or other content words.
- If γ_j is medium, the partial translation has a larger impact on the prediction of y_j, which is slightly related to x. Here, y_j may be prepositions, determiners, or other function words.
- If γ_j is small, the model relies on the language model to produce the unrelated translations, which are exactly our targeted noisy samples.

3.2 Norm-based Corpus Filtering

Based on the metric γ_j calculated at each step, we measure how much the target sentence y is aligned with the input x as follows:

$$\mathcal{R}(\mathrm{x}, \mathrm{y}) = \frac{1}{m} \sum_{j=1}^{m} \gamma_j \tag{6}$$

When $\mathcal{R}(\mathrm{x}, \mathrm{y})$ is smaller than a threshold k, the target sentence may be desperately inadequate or even wholly unrelated to the source sentence. To eliminate the impact of these noisy samples, we erase their loss during training by the norm-based sentence-level objective:

$$\mathcal{L} = I_{\mathcal{R}(\mathrm{x},\mathrm{y})>k} \cdot \mathcal{L}_{\mathrm{NLL}} \tag{7}$$

where the indicative function $I_{\mathcal{R}(\mathrm{x},\mathrm{y})>k}$ is equal to 1 if $\mathcal{R}(\mathrm{x}, \mathrm{y}) > k$, else 0. k is a hyperparameter which is used to adjust the filtering ratio. $\mathcal{L}_{\mathrm{NLL}}$ is the loss of the NMT calculated by the negative log-likelihood in Equation (1).

Considering the early NMT model is not well-trained to gather information from the source and target sides, we first warm up the model on the entire dataset for T steps and then filter the noisy sentence pairs based on observed model behaviors. In the middle and later stages, $I_{\mathcal{R}(\mathrm{x},\mathrm{y})>k}$ would stable at a particular value.

3.3 Noisy Label Refurbishing

The detected unparallel sentence pairs hamper the training of the NMT system. But they can split into individual monolingual sentences, which remain to be fully re-utilized. From this perspective, we propose to refurbish these noisy samples by generating pseudo labels via knowledge distillation (Hinton et al., 2015; Kim and Rush, 2016).

The biggest issue of integrating knowledge distillation in our scenario is how to acquire a strong teacher model, which decides the performance of the student model (Gou et al., 2021). However, the absence of a large-scale clean corpus makes it hard to train an offline competitive teacher model. Alternatively, we employ online self-distillation (Wei et al., 2019) to let the history model generate the translation for noisy samples.

Concretely, we use the checkpoint with the best performance on the validation set as the teacher. Then, the current model learns to match the teacher model's prediction $q(\cdot|x)$ on the noisy data. The word-level self-distillation loss can be defined as:

$$\mathcal{L}_{\mathrm{SD}} = -\sum_{j=1}^{m} \sum_{i=1}^{|\mathcal{V}|} q(y_j = i|\mathrm{y}_{<j}, \mathrm{x}; \theta_{\mathcal{T}}) \times \log p(y_j = i|\mathrm{y}_{<j}, \mathrm{x}; \theta) \tag{8}$$

where $\theta_{\mathcal{T}}$ and θ parameterize the teacher and student model separately. $\mathcal{V}$ is the target vocabulary set. In this way, we make full use of the corpus by precisely figuring out the clean data and replacing the remaining noises with pseudo labels:

$$\mathcal{L} = I_{\mathcal{R}(\mathrm{x},\mathrm{y})>k} \cdot \mathcal{L}_{\mathrm{NLL}} + I_{\mathcal{R}(\mathrm{x},\mathrm{y})\leq k} \cdot \mathcal{L}_{\mathrm{SD}} \tag{9}$$

4 Experiments

We conduct experiments on two types of datasets: (1) WMT 2020 shared task on parallel corpus fil-

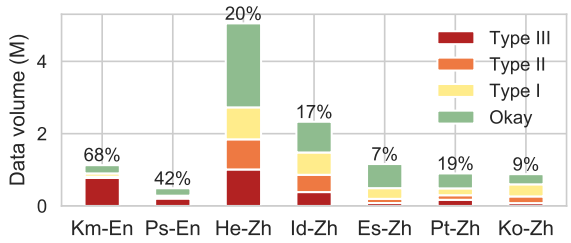

Figure 5: Statistics of datasets after rule-based filtering. We annotate the proportion of type III noises in each dataset. Our noisy corpora filtering method targets type III and partial type II.

tering and alignment for low-resource conditions[2]: Khmer⇒English (Km⇒En) and Pashto⇒English (Ps⇒En), and (2) our in-house web crawled corpora: Hebrew (He), Indonesian (Id), Korean (Ko), Portuguese (Pt), and Spanish (Es) to Chinese (Zh).

4.1 Dataset

WMT20 corpus filtering task asks participants to select different-scale subsets of high-quality sentence pairs from the noisy data. The quality of selected subsets is measured by the performance of an NMT system trained on this data. This task provides three kinds of data: (1) 4.17M Km⇒En and 1.02M Ps⇒En noisy sentence pairs which participants have to score for filtering, and (2) clean parallel and monolingual data to train quality estimation models that help the filtering task, and (3) development and test sets used to evaluate translation systems trained on filtered data. Note that we do not use the second part of the data and only experiment with (1) and (3). We strictly follow Koehn et al. (2020) to preprocess the raw data. For the in-house corpora, we apply sentence pieces on tokenized text. We construct the vocabulary with the size of 20k tokens where the source and target languages are separately encoded.

The raw corpora are firstly filtered by heuristic rules to remove extremely noisy sentence pairs. We implement rule-based filtering as in Lu et al. (2020), the details of which are listed in Appendix B.

A cursory review of the above corpora is given in Figure 5. We categorize unparallel sentence pairs into three types based on the level of misalignment: (I) words, (II) phrases, and (III) the whole sentence. We randomly sample 200 sentence pairs from each preprocessed dataset and manually annotate them with predefined labels based on their noise degrees.

[2]https://www.statmt.org/wmt20/parallel-corpus-filtering.html

We find a high noise rate in the WMT20 corpora, while noise types in in-house datasets are diverse. These two kinds of datasets pose different challenges for our methods to filter noises accurately.

4.2 Settings

We strictly follow model configurations and evaluation settings provided by WMT20 organizers. The evaluation is done on subsets of two predefined sizes, 5M and 7M English words. The most striking difference between participants and us is that we simultaneously perform data filtering and model training rather than "filter first and next train".

For our in-house datasets, we experiment with Transformer `Base` (Vaswani et al., 2017). More details about experimental settings for WMT20 and in-house datasets are given in Appendix C.

The choice of the threshold k is key to our methods. In practice, we rank 200 samples manually labeled with noise extents in Section 4.1 by $\mathcal{R}(\mathrm{x}, \mathrm{y})$. k is set based on different scenarios. If given the remaining data size, i.e., WMT20 predefines the size of selected data, we select the corresponding k in 200 annotated samples by the ratio of remaining data. In the WMT20 scenario, k is 2.75 and 2.45 for the 5M and 7M words setting in Km⇒En. For Ps⇒En, k is 2.3 and 1.4 for those two settings. In the case of no required size for the data left, we set k based on the noise rate of annotated samples and filter the noisiest samples ranking at the bottom. For various noise rates in in-house datasets, k is 1.8 for He, Pt, Ko⇒Zh, 1.65 for Id⇒Zh, and 1.9 for Es⇒Zh. We find that the model capacity affects the value of $\mathcal{R}(\mathrm{x}, \mathrm{y})$, making k differ greatly for experiments on WMT20 and in-house datasets (model parameters 47M vs. 68M). Besides, it is easy to see that different language pairs have similar ranges of $\mathcal{R}(\mathrm{x}, \mathrm{y})$ under one experimental setting.

4.3 Main Results

To thoroughly compare with participants in the WMT20 corpus filtering task, we report the performance of two models ranking the top (Esplà-Gomis et al., 2020; Lu et al., 2020) and the official baseline LASER (Artetxe and Schwenk, 2019). The best showings leverage the clean external parallel and monolingual data to score each language pair, whereas we do not use this part of the data.

Table 1 presents the performance of the NMT model trained on participants' selected subsets with varying scales. Our proposed method yields comparable results with the best results in this competi-

Methods	Khmer⇒English		Pashto⇒English	
	DEVT	TEST	DEVT	TEST
Raw Data	1.1	1.3	6.6	4.3
+ Rule-based pre-filtering	4.2	4.5	9.7	7.4
LASER (Artetxe and Schwenk, 2019)*	7.1 / 6.7	8.4 / 8.6	9.7 / 9.7	7.7 / 8.2
UA-Prompsit (Esplà-Gomis et al., 2020)*	8.4 / 7.6	10.0 / 9.4	10.8 / 10.2	9.2 / 8.4
Alibaba (Lu et al., 2020)*	8.9 / 7.8	11.0 / 10.1	10.8 / 10.0	9.5 / 8.8
Norm-based corpus filering	8.3 / 7.5	9.8 / 9.3	10.9 / 10.4	9.5 / 8.4

Table 1: Translation results for WMT20 parallel corpus filtering task. BLEU scores are reported for systems trained on subsets of the data (5M / 7M words), subsampled based on different quality scores. LASER scores are an officially provided baseline. The best showing of this competition is by Alibaba, followed by UA-Prompsit. Other submissions are at least 0.5 BLEU points behind these. * denotes that the results come from the cited paper.

Methods	He⇒Zh	Id⇒Zh	Pt⇒Zh	Ko⇒Zh	Es⇒Zh	AVE	Δ
Raw Data	14.51	44.16	10.75	12.01	12.35	18.01	–
+ Rule-based pre-filtering	16.50	45.38	10.98	11.96	12.47	18.44	+ 0.34
+ Norm-based corpus filering	16.66	45.93	11.25	12.58	12.61	18.71	+ 0.70
+ Noisy label refurbishing	16.82	46.61	12.60	13.22	12.89	19.12	+ 1.11

Table 2: Evaluation of translation quality for our in-house corpora using case-insensitive BLEU scores.

tion. It performs as well as the strongest competitor in Ps⇒En and takes second place in Km⇒En, showing our method's effectiveness in identifying noisy samples. Compared with them, our approach has two main advantages: (1) no need for clean parallel data, which is unavailable for low-resource languages, and (2) low time and computation costs achieved by incorporating data filtering and model training into one process. We take Lu et al. (2020) as an example. They ensemble eight models to score each sentence pair, including dual bilingual GPT-2 models (Radford et al., 2019), dual conditional cross-entropy models (Junczys-Dowmunt, 2018), IBM word alignment models of two directions (Khadivi and Ney, 2005), and GPT-2 language models of the source and target side. The cost of training those models is high, while similar costs have been also reported in other submissions.

Translation results on our in-house datasets are shown in Table 2. The norm-based corpora filtering outperforms the baseline by 2.15 (He⇒Zh), 1.77 (Id⇒Zh), 0.50 (Pt⇒Zh), 0.57 (Ko⇒Zh), and 0.26 (Es⇒Zh), respectively. It also exceeds the rule-based filtering method. Moreover, based on discovered unparallel sentence pairs, our noisy label refurbishing method yields improvements of 2.31, 2.45, 1.85, 1.21, and 0.54 BLEU scores on He, Id, Pt, Ko, and Es⇒Zh. Besides, performing online knowledge distillation only adds 9% training time to the baseline. It is acceptable concerning performance gains obtained in this process.

We find that further benefits from our methods vary across different datasets, which are minor in He⇒Zh and Es⇒Zh but extremely significant in Id⇒Zh and Pt⇒Zh. There are two main reasons for that: the scale of the dataset and the noise rate. A large-scale dataset in He⇒Zh makes it robust to a high percentage of noises (Jayanthi and Pratapa, 2021). On the other hand, as seen in Figure 5, the type III noise, which presents the most misaligned sentence pairs, only accounts for 7% in Es⇒Zh, which leads to low demand for noisy data filtering.

Notably, our method has a specific scope of applications. We do not suggest using noisy label refurbishing when the noise rate[3] exceeds 30%, i.e., WMT20 datasets, for massive noises lead to a weak baseline model. However, our norm-based corpora filtering method still works in these cases.

Changes of $I_{\mathcal{R}(\mathrm{x},\mathrm{y})\leq k}$ During Training. $\mathcal{R}(\mathrm{x},\mathrm{y})$ is a dynamic indicator based on current model behaviors. Figure 6 exhibits changes of the proportion of $I_{\mathcal{R}(\mathrm{x},\mathrm{y})\leq k}$ and the accuracy of our proposed methods. As the training progresses, our model becomes more competent with richer knowledge from the teacher involved. The percentage of $I_{\mathcal{R}(\mathrm{x},\mathrm{y})\leq k}$ increases rapidly at first and then flattens out, which proves the stability of our metric.

[3]The ratio of sentence-level unparallel pairs. The criteria of category is introduced in section 4.1.

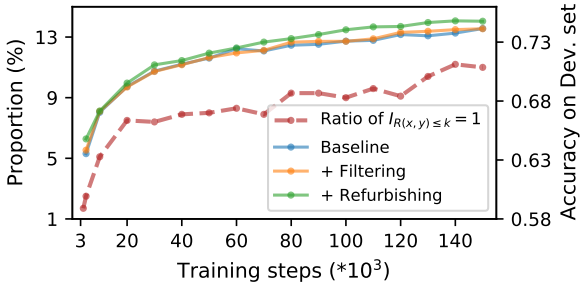

Figure 6: Changes of the proportion of $I_{\mathcal{R}(\mathrm{x},\mathrm{y})\leq k}$ on Id⇒Zh during training and accuracy on the validation set. We warm up the model for $T = 3k$ steps and then filter samples where $I_{\mathcal{R}(\mathrm{x},\mathrm{y})\leq k} = 1$. The red line refers to the left y-axis, and others refer to the right y-axis.

Methods		He⇒Zh
Ours		16.82
Teacher model	Last	16.52
Selective distillation	All	16.06
	Half (rank-high)	15.87
	Half (rank-low)	16.62
Distillation mode	FT (sample)	15.76
	FT (beam search)	16.56

Table 3: Comparisons of different knowledge distillation in He⇒Zh from three perspectives. The first two types are both online distillation. "Last" means using the last checkpoint as the teacher model. Selective distillation is to choose different partition of distilled samples: (1) All: the whole training set, (2) Half (rank-high): 50% top of $\mathcal{R}(\mathrm{x},\mathrm{y})$, and (3) Half (rank-low): 50% bottom of $\mathcal{R}(\mathrm{x},\mathrm{y})$. The third is offline distillation, Forward Translation (Zhang and Zong, 2016). "sample" and "beam search" are two inference ways to get the synthetic data.

Variations of Knowledge Distillation. As aforesaid, we use the best checkpoint on the validation set as the teacher model to distill located noisy samples only. It raises the question whether we have better options for the teacher model or whether we can conduct a wide range of knowledge distillation. For comparison, we try more variations of knowledge distillation and present results in Table 3. We find that the best checkpoint is more competent than the last one, which is largely similar to the current model with limited complementary knowledge to the student. For selective distillation, we see that distilling the whole dataset is not a good choice. The amount of teacher knowledge is not "more is better" (Wang et al., 2021). It may induce more noise, especially for a weak teacher model. Among them, the bottom 50% of $\mathcal{R}(\mathrm{x},\mathrm{y})$ are in a higher demand for distillation. Those samples with

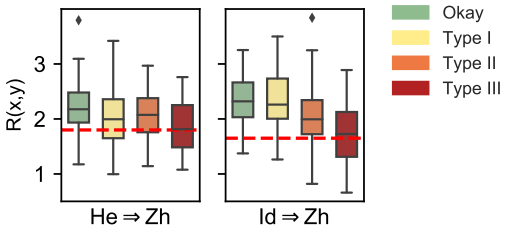

Figure 7: $\mathcal{R}(\mathrm{x},\mathrm{y})$ of manually annotated samples with varying noise degrees in section 4.1. Type I, II, and III represent word-, phrase-, and sentence-level misalignment, respectively. Type III and partial type II are our target noisy samples that need to be filtered. The dashed red line is the threshold k we set.

higher $\mathcal{R}(\mathrm{x},\mathrm{y})$ are likely with clean labels where the teacher model is useless. The results show the necessity of carefully selecting distilled samples in the presence of noise.

Furthermore, our method is related to Forward Translation (FT) (Zhang and Zong, 2016) for exploiting the monolingual data. They use the earlier trained model as the teacher to translate source sentences to target translation, and the obtained synthetic corpora are fed to the student model trained later. To study the usefulness of FT in our scenario, we regard our proposed noisy data filtering method as the teacher. Then, we split misaligned samples, where $\mathcal{R}(\mathrm{x},\mathrm{y}) \leq k$ (1.01M in He⇒Zh), as monolingual source sentences. The following steps are in line with FT. From the last two lines in Table 3, the synthetic data is of poor quality if generated by sampling for a weak teacher model. Beam search ensures good translations and performs better but is computationally expensive. Unlike sentence-level distillation, the word-level distillation in this paper allows the transfer of local word distributions. It eliminates the error propagated from the teacher model, which is more suitable in the noise scenario.

5 Analysis

We conduct extensive analyses to evaluate the ability of $\mathcal{R}(\mathrm{x},\mathrm{y})$ to pinpoint unparallel sentence pairs. We first examine whether $\mathcal{R}(\mathrm{x},\mathrm{y})$ can reflect the overall degree of misalignment. From a more fine-grained view, we explore the correlation between the score γ_j at step j and linguistic properties.

5.1 Correlation with the Misalignment Degree

As previous, we filter sentence pairs where $\mathcal{R}(\mathrm{x},\mathrm{y})$ is lower than the threshold k. To explore whether filtered samples are indeed corrupted, we calculate $\mathcal{R}(\mathrm{x},\mathrm{y})$ of annotated samples in section 4.1,

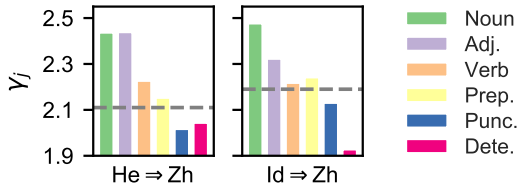

Figure 8: γ_j of target-side words with different POS tags in He⇒Zh and Id⇒Zh. The dashed gray line is the average score.

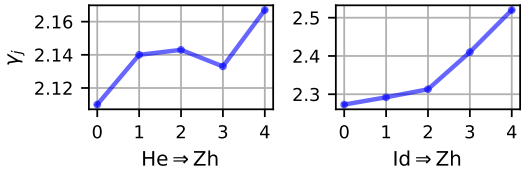

Figure 9: γ_j versus the fertility of the target-side word in He⇒Zh and Id⇒Zh.

which are categorized into four classes based on the degree of misalignment. As shown in Figure 7, $\mathcal{R}(\mathrm{x}, \mathrm{y})$ is negatively correlated with the extent of noises. Type III and part of type II noisy samples are assigned with relatively lower $\mathcal{R}(\mathrm{x}, \mathrm{y})$ where too much target-side information is required to predict the translation. It indicates that our proposed measurement is reflective of misalignment and sufficient to filter unparallel sentence pairs for NMT.

5.2 Correlation with Linguistic Properties

As seen in Equation (6), the sentence-level $\mathcal{R}(\mathrm{x}, \mathrm{y})$ is averaged over γ_j at step j, which depicts whether each target word corresponds to any source words. This section studies the relation between γ_j and two properties, syntactic roles and fertility. Chinese sentences are POS tagged by `jieba`[4]. Fertility reveals how many source tokens a target token is aligned to, which is obtained by fast align (Dyer et al., 2013) to extract bilingual alignment. Results are reported on the validation set.

As shown in Figure 8, content words are in great need of the source context, thus leading to a higher γ_j. However, content-free words, like punctuation and determiner, mainly rely on the target-side information, where γ_j is significantly below average. Furthermore, the value of γ_j is positively related to the fertility of the target word. As illustrated in Figure 9, the prediction of the target token aligning to more source words relies more on the source sentence, thus leading to a higher γ_j. These findings fully show the rationality of our proposed metrics.

[4]https://pypi.org/project/jieba/

6 Related Work

Many web-crawled data for training the NMT system are so noisy that we should select the high-quality subset. In this section, we first review recent advances in noisy corpora filtering for NMT. As we treat the discovered noisy data as unlabeled monolingual data to distill in this paper, another related work is knowledge distillation in NMT.

6.1 Parallel Corpus Filtering

There is a rich body of work on filtering out noises in parallel data. Xu and Koehn (2017) construct the noisy synthetic data (inadequate and non-fluent translations) and train a classifier to distinguish good from the bad. Açarçiçek et al. (2020) follow this idea with a classifier based on a multilingual version of the RoBERTa (Conneau et al., 2020). Many other studies employ different bilingual and monolingual language models to score the sentence pairs (Lu et al., 2020; Esplà-Gomis et al., 2020), which are data-hungry and time-consuming in practice. Unlike them, our method does not use external data and yields comparable results. Also, we perform data filtering and model training in one stage to reduce time and computation costs.

6.2 Knowledge Distillation in NMT

Knowledge distillation transfers the knowledge from the teacher to the student model. It is widely studied in NMT to obtain a lightweight and effective model. Kim and Rush (2016) use an offline large-capacity NMT system as the fixed teacher model, which is also extended to multilingual NMT (Tan et al., 2019). Instead, the same-capacity teacher model is used in Zhang and Zong (2016); Sennrich et al. (2016). Such approaches need massive clean data for training an accurate teacher model, which is impractical in the limited noisy corpora. Another line of work applies self-distillation to NMT using the current or history model as the teacher model (Wei et al., 2019; Hahn and Choi, 2019), updating the distilled knowledge as a better model comes. Here, we focus on a new scenario where self-distillation is employed to relabel the noisy samples in the training of the noisy corpora.

7 Conclusion

This paper presents a novel norm-based noisy corpora filtering and refurbishing method. We propose to use the information ratio from the source to the target side to distinguish unparallel sentence pairs.

The amounts of these two information flows are calculated by norms of context vectors of each side. Unlike parallel sentence pairs, the excessive target-side information is needed for the model to fit unparallel ones, which present relatively lower scores. We incorporate the noisy corpora filtering into the training of NMT without any extra clean data or costly pre-trained scorers. Extensive experiments show that our method performs comparably with SOTA results with significant advantages in time and computational costs. We further refurbish the discovered noisy data by producing pseudo labels via online knowledge distillation, which obtains further performance gains.

8 Limitations

Our methods have a specific scope of applications due to the methodology design. As highlighted in Section 3, the basis of our approach is the difference in how the model processes unparallel and parallel sentence pairs. Thus, it cannot work well when a large extent of noise makes the NMT model hard to converge. We take Nepali-English in WMT 2019 shared task on parallel corpus filtering and alignment[5] as an example. By sampling inspection, we find that 89.4% of the dataset is wholly misaligned after pre-processing. It is unstable to directly train an NMT model with the whole dataset, which results in our worse performance than the best showing (2.2 BLEU vs. 3.2 BLEU). In this case, extra clean data or resources of similar languages are necessary to build a competent scorer.

Acknowledgements

This work is supported by the Natural Science Foundation of China under Grant 62122088 and U1836221.

References

Haluk Açarçiçek, Talha Çolakoğlu, Pınar Ece Aktan Hatipoğlu, Chong Hsuan Huang, and Wei Peng. 2020. Filtering noisy parallel corpus using transformers with proxy task learning. In *Proceedings of the Fifth Conference on Machine Translation*, pages 940–946, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

[5]https://www.statmt.org/wmt19/parallel-corpus-filtering.html

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*, abs/1607.06450.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, pages 644–648.

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, and Felipe Sánchez-Martínez. 2020. Bicleaner at WMT 2020: Universitat d'alacant-prompsit's submission to the parallel corpus filtering shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 952–958, Online. Association for Computational Linguistics.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *Int. J. Comput. Vis.*, 129(6):1789–1819.

Sangchul Hahn and Heeyoul Choi. 2019. Self-knowledge distillation in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 423–430, Varna, Bulgaria. INCOMA Ltd.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

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

Sai Muralidhar Jayanthi and Adithya Pratapa. 2021. A study of morphological robustness of neural machine translation. *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Shahram Khadivi and Hermann Ney. 2005. Automatic filtering of bilingual corpora for statistical machine translation. In *Natural Language Processing and Information Systems, 10th International Conference on Applications of Natural Language to Information*

Systems, NLDB 2005, Alicante, Spain, June 15-17, 2005, Proceedings, volume 3513 of *Lecture Notes in Computer Science*, pages 263–274. Springer.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. 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.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online. Association for Computational Linguistics.

Jun Lu, Xin Ge, Yangbin Shi, and Yuqi Zhang. 2020. Alibaba submission to the WMT20 parallel corpus filtering task. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 979–984. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Open AI blog*.

Adriaan M. J. Schakel and Benjamin J. Wilson. 2015. Measuring word significance using distributed representations of words. *ArXiv*, abs/1508.02297.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2214–2218. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010. Curran Associates Inc.

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

Hao-Ran Wei, Shujian Huang, Ran Wang, Xin-yu Dai, and Jiajun Chen. 2019. 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)*, pages 1932–1941, Minneapolis, Minnesota. Association for Computational Linguistics.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Copenhagen, Denmark. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

# Sentence pairs	WMT 2020		In-house datasets				
	Km-En	Ps-En	He-Zh	Id-Zh	Es-Zh	Pt-Zh	Ko-Zh
Raw data	4.17M (58.35M)	1.02M (11.55M)	5.72M	3.17M	1.50M	1.10M	1.00M
+ Rule-based pre-filtering	1.13M (20.27M)	0.49M (7.78M)	5.05M	2.49M	1.16M	0.90M	0.88M

Table 4: Statistics of preprocessed data. We list the number of English words in parentheses in WMT20 datasets.

A Norm-based Word Importance Measurement

We present the relation between the norm of c_i^{L-1} and token frequency in Figure 10. We can see that the norm of c_i^{L-1} decreases with a high word frequency. Moreover, the distribution of blue dots (function words) is lower than that of green dots (content words). It indicates that words with more diverse semantics receive higher norms. The observations in Figure 3 and 10 show that we can infer how much information is encoded in word representations from the perspective of norms.

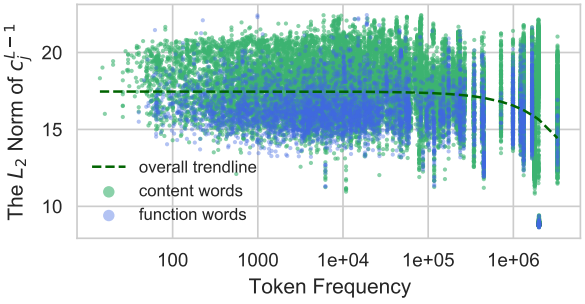

Figure 10: The L_2 norm of c_j^{L-1} versus token frequency of English words in the LDC Chinese-to-English vocabulary labeled content words (green dots) and function words (blue dots). This produces a downward trendline.

B Rule-based Pre-filtering

Following Lu et al. (2020), we apply a series of heuristic rules to filter the low-quality sentence pairs, which includes:

- *The length of the sentence.* The too short ($\leq$ 2 words) or too long ($>$ 50 words) sentences will be removed.
- *The length ratio of the source sentence to the target sentence.* The ratio is set between 0.2 to 5 for all language pairs.
- *The proportion of valid tokens.* A valid token should include the letters in the corresponding language. The sentence is dropped if the valid-token ratio is less than 0.2.
- *Language filtering.* We detect the language of a sentence by using a language detection tool *fasttext*. It helps remove the sentence pairs if either the source or the target sentence does not belong to the required language.
- *URLs or numbers.* We remove the sentence which contains URLs or more than 25% numerical tokens.

The size of data after ruled-based pre-filtering is given in Table 4. We find that WMT20 datasets are very noisy, where around 72.66% of Km-En and 51.96% of Ps-En are filtered out. The noise rate in He-Zh is relatively lower. As shown in Table 1, pre-filtering is necessary to relieve stress for the following filtering method.

C Experimental Settings

For WMT20 datasets, we strictly follow the model configuration and evaluation settings provided by the WMT20 organizers (Koehn et al., 2020). It includes five stacked encoder layers and five stacked decoder layers. During training, we use Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.998$, an inverse sqrt learning rate of 4,000 warm-up steps, and dropout is 0.4. All experiments last for 100 epochs with a single GPU, where the batch size is 4000 tokens. We accumulate the gradient of parameters and update every 4 steps. Scores on test sets are reported by case-insensitive Sacrebleu (Post, 2018).

For our in-house datasets, we experiment with Transformer `Base` (Vaswani et al., 2017). During training, we use label smoothing of value $\epsilon_{ls} = 0.1$ and employ the Adam ($\beta_1 = 0.9, \beta_2 = 0.998$) for parameter optimization with a scheduled learning rate of 4,000 warm-up steps. The training lasts 150k for He⇒Zh and Id⇒Zh, 100k for the other three. We average the last ten checkpoints and use beam search (beam size 5, length penalty 1.2) for inference. We measure case-insensitive BLEU calculated by *multi-bleu.perl*.

Our method introduces two hyper-parameters. The first one is the warm-up step T to ensure the

stability of our metric, and the second is the threshold k to determine the amount of the filtered data. For the former, we observe that our proposed metric reaches a stable range after several thousand steps and thus set $T = 3k$. As aforesaid in Section 4.2, we determine the threshold k based on the percentage of the remaining data if given the required size of selected subsets. We take Ps-En as an example. The dataset after rule-based filtering contains 7.78M English words. Thus, we should remove 10% and 35% of the data to obtain the 5M and 7M words settings. We determine the threshold k as the lower decile in the $\mathcal{R}(\mathrm{x}, \mathrm{y})$ of 200 annotated samples for 10% removal, similar to 35%.