Self-Training Sampling with Monolingual Data Uncertainty for Neural Machine Translation

Wenxiang Jiao[†*] **Xing Wang**[‡] **Zhaopeng Tu**[‡] **Shuming Shi**[‡] **Michael R. Lyu**[†] **Irwin King**[†]
[†]Department of Computer Science and Engineering
The Chinese University of Hong Kong, HKSAR, China
[‡]Tencent AI Lab
[†]{wxjiao,lyu,king}@cse.cuhk.edu.hk
[‡]{brightxwang,zptu,shumingshi}@tencent.com

Abstract

Self-training has proven effective for improving NMT performance by augmenting model training with synthetic parallel data. The common practice is to construct synthetic data based on a randomly sampled subset of large-scale monolingual data, which we empirically show is sub-optimal. In this work, we propose to improve the sampling procedure by selecting the most informative monolingual sentences to complement the parallel data. To this end, we compute the uncertainty of monolingual sentences using the bilingual dictionary extracted from the parallel data. Intuitively, monolingual sentences with lower uncertainty generally correspond to easy-to-translate patterns which may not provide additional gains. Accordingly, we design an uncertainty-based sampling strategy to efficiently exploit the monolingual data for self-training, in which monolingual sentences with higher uncertainty would be sampled with higher probability. Experimental results on large-scale WMT English⇒German and English⇒Chinese datasets demonstrate the effectiveness of the proposed approach. Extensive analyses suggest that emphasizing the learning on uncertain monolingual sentences by our approach does improve the translation quality of high-uncertainty sentences and also benefits the prediction of low-frequency words at the target side.[1]

1 Introduction

Leveraging large-scale unlabeled data has become an effective approach for improving the performance of natural language processing (NLP) models (Devlin et al., 2019; Brown et al., 2020; Jiao et al., 2020a). As for neural machine translation (NMT), compared to the parallel data, the monolingual data is available in large quantities for many languages. Several approaches on boosting the NMT performance with the monolingual data have been proposed, e.g., data augmentation (Sennrich et al., 2016a; Zhang and Zong, 2016), semi-supervised training (Cheng et al., 2016; Zhang et al., 2018; Cai et al., 2021), pre-training (Siddhant et al., 2020; Liu et al., 2020). Among them, data augmentation with the synthetic parallel data (Sennrich et al., 2016a; Edunov et al., 2018) is the most widely used approach due to its simple and effective implementation. It has been a de-facto standard in developing the large-scale NMT systems (Hassan et al., 2018; Ng et al., 2019; Wu et al., 2020; Huang et al., 2021).

Self-training (Zhang and Zong, 2016) is one of the most commonly used approaches for data augmentation. Generally, self-training is performed in three steps: (1) randomly sample a subset from the large-scale monolingual data; (2) use a "teacher" NMT model to translate the subset data into the target language to construct the synthetic parallel data; (3) combine the synthetic and authentic parallel data to train a "student" NMT model. Recent studies have shown that synthetic data manipulation (Edunov et al., 2018; Caswell et al., 2019) and training strategy optimization (Wu et al., 2019b; Wang et al., 2019) in the last two steps can boost the self-training performance significantly. However, how to efficiently and effectively sample the subset from the large-scale monolingual data in the first step has not been well studied.

Intuitively, self-training simplifies the complexity of generated target sentences (Kim and Rush, 2016; Zhou et al., 2019; Jiao et al., 2020b), and easy patterns in monolingual sentences with deterministic translations may not provide additional gains over the self-training "teacher" model (Shrivastava et al., 2016). Related work on computer

[*] Work was mainly done when Wenxiang Jiao was interning at Tencent AI Lab.

[1] The source code is available at https://github.com/wxjiao/UncSamp

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2840–2850
August 1–6, 2021. ©2021 Association for Computational Linguistics

vision also reveals that easy patterns in unlabeled data with the deterministic prediction may not provide additional gains (Mukherjee and Awadallah, 2020). In this work, we investigate and identify the uncertain monolingual sentences which implicitly hold difficult patterns and exploit them to boost the self-training performance. Specifically, we measure the uncertainty of the monolingual sentences by using a bilingual dictionary extracted from the authentic parallel data (§2.1). Experimental results show that NMT models benefit more from the monolingual sentences with higher uncertainty, except on those with excessively high uncertainty (§2.3). By conducting the linguistic property analysis, we find that extremely uncertain sentences contain relatively poor translation outputs, which may hinder the training of NMT models (§2.4).

Inspired by the above finding, we propose an uncertainty-based sampling strategy for self-training, in which monolingual sentences with higher uncertainty would be selected with higher probability (§3.1). Large-scale experiments on WMT English⇒German and English⇒Chinese datasets show that self-training with the proposed uncertainty-based sampling strategy significantly outperforms that with random sampling (§3.3). Extensive analyses on the generated outputs confirm our claim by showing that our approach improves the translation of uncertain sentences and the prediction of low-frequency target words (§3.4).

Contributions. Our main contributions are:

- We demonstrate the necessity of distinguishing monolingual sentences for self-training.
- We propose an uncertainty-based sampling strategy for self-training, which selects more complementary sentences for the authentic parallel data.
- We show that NMT models benefit more from uncertain monolingual sentences in self-training, which improves the translation quality of uncertain sentences and the prediction accuracy of low-frequency words.

2 Observing Monolingual Uncertainty

In this section, we aimed to understand the effect of uncertain monolingual data on self-training. We first introduced the metric for identifying uncertain monolingual sentences, then the experimental setup and at last our preliminary results.

Notations. Let X and Y denote the source and target languages, and let $\mathcal{X}$ and $\mathcal{Y}$ represent the sentence domains of corresponding languages. Let $\mathcal{B} = \{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^N$ denote the authentic parallel data, where $\mathbf{x}^i \in \mathcal{X}$, $\mathbf{y}^i \in \mathcal{Y}$ and N is the number of sentence pairs. Let $\mathcal{M}_x = \{\mathbf{x}^j\}_{j=1}^{M_x}$ denote the collection of monolingual sentences in the source language, where $\mathbf{x}^j \in \mathcal{X}$ and M_x is the size of the set. Our objective is to obtain a translation model $f : \mathcal{X} \mapsto \mathcal{Y}$, that can translate sentences from language X to language Y.

2.1 Identification of Uncertain Data

Data Complexity. According to Zhou et al. (2019), the complexity of a parallel corpus can be measured by adding up the translation uncertainty of all source sentences. Formally, the translation uncertainty of a source sentence $\mathbf{x}$ with its translation candidates can be operationalized as conditional entropy:

$$\mathcal{H}(\mathbf{Y}|\mathbf{X}=\mathbf{x}) = -\sum_{\mathbf{y}\in\mathcal{Y}} p(\mathbf{y}|\mathbf{x}) \log p(\mathbf{y}|\mathbf{x}) \quad (1)$$

$$\approx \sum_{t=1}^{T_x} \mathcal{H}(y|x=x_t), \quad (2)$$

where T_x denotes the length of the source sentence, x and y represent a word in the source and target vocabularies, respectively. Generally, a high $\mathcal{H}(\mathbf{Y}|\mathbf{X}=\mathbf{x})$ denotes that a source sentence $\mathbf{x}$ would have more possible translation candidates.

Equation (2) estimates the translation uncertainty of a source sentence with all possible translation candidates in the parallel corpus. It can not be directly applied to the sentences in monolingual data due to the lack of corresponding translation candidates. One potential solution to the problem is utilizing a trained model to generate multiple translation candidates. However, generation may lead to bias estimation due to the generation diversity issue (Li et al., 2016; Shu et al., 2019). More importantly, generation is extremely time-consuming for large-scale monolingual data.

Monolingual Uncertainty. To address the problem, we modified Equation (2) to reflect the uncertainty of monolingual sentences. We estimate the target word distribution conditioned on each source word based on the authentic parallel corpus, and then use the distribution to measure the translation uncertainty of the monolingual example. Specifically, we measure the uncertainty of monolingual sentences based on the bilingual dictionary.

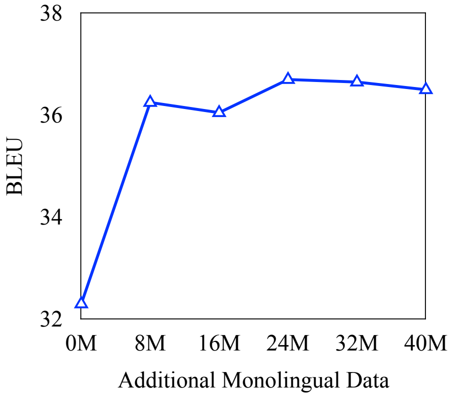


Figure 1: Performance of self-training with increased size of monolingual data. The BLEU score is averaged on WMT En⇒De newstest2019 and newstest2020.

For a given monolingual sentence $\mathbf{x}^j \in \mathcal{M}_x$, its uncertainty U is calculated as:

$$\mathrm{U}(\mathbf{x}^j|\mathcal{A}_b) = \frac{1}{T_x}\sum_{t=1}^{T_x}\mathcal{H}(y|\mathcal{A}_b, x = x_t), \qquad (3)$$

which is normalized by T_x to avoid the length bias. A higher value of U indicates a higher translation uncertainty of the monolingual sentence.

In Equation 3, the word level entropy $\mathcal{H}(y|\mathcal{A}_b, x = x_t)$ captures the translation modalities of each source word by using the bilingual dictionary $\mathcal{A}_b$. The bilingual dictionary records all the possible target words for each source word, as well as translation probabilities. It can be built from the word alignments by external alignment toolkits on the authentic parallel corpus. For example, given a source word x with all three word translations y_1, y_2 and y_3 and the translation probabilities of $p(y_1|x)$, $p(y_2|x)$ and $p(y_3|x)$, respectively, the word level entropy can be calculated as follows:

$$\mathcal{H}(y|\mathcal{A}_b, x_i) = -\sum_{y_j \in \mathcal{A}_b(x_i)} p(y_j|x_i)\log p(y_j|x_i). \qquad (4)$$

2.2 Experimental Setup

Data. We conducted experiments on two large-scale benchmark translation datasets, i.e., WMT English⇒German (En⇒De) and WMT English⇒Chinese (En⇒Zh). The authentic parallel data for the two tasks consists of about 36.8M and 22.1M sentence pairs, respectively. The monolingual data we used is from newscrawl released by WMT2020. We combined the newscrawl data from year 2011 to 2019 for the English monolingual corpus, consisting of about 200M sentences. We randomly sampled 40M monolingual data for En⇒De and 20M for En⇒Zh unless otherwise stated. We adopted newstest2018 as the validation set and used newstest2019/2020 as the test sets. For each language pair, we applied Byte Pair Encoding (BPE, Sennrich et al., 2016b) with 32K merge operations.

Model. We chose the state-of-the-art TRANSFORMER (Vaswani et al., 2017) network as our model, which consists of an encoder of 6 layers and a decoder of 6 layers. We adopted the open-source toolkit Fairseq (Ott et al., 2019) to implement the model. We used the TRANSFORMER-BASE model for preliminary experiments (§2.3) and the constrained scenario (§3.2) for efficiency. For the unconstrained scenario (§3.3), we adopted the TRANSFORMER-BIG model. Results on these models with different capacities can also reflect the robustness of our approach. For the TRANSFORMER-BASE model, we trained it for 150K steps with 32K (4096 × 8) tokens per batch. For the TRANSFORMER-BIG model, we trained it for 30K steps with 460K (3600 × 128) tokens per batch with the cosine learning rate schedule (Wu et al., 2019a). We used 16 Nvidia V100 GPUs to conduct the experiments and selected the final model by the best perplexity on the validation set.

Evaluation. We evaluated the models by BLEU score (Papineni et al., 2002) computed by SacreBLEU (Post, 2018)[2]. For the En⇒Zh task, we added the option `--tok zh` to SacreBLEU. We measured the statistical significance of improvement with paired bootstrap resampling (Koehn, 2004) using `compare-mt`[3] (Neubig et al., 2019).

2.3 Effect of Uncertain Data

First of all, we investigated the effect of monolingual data uncertainty on the self-training performance in NMT. We conducted the preliminary experiments on the WMT En⇒De dataset with the TRANSFORMER-BASE model. We sampled 8M bilingual sentence pairs from the authentic parallel data and randomly sampled 40M monolingual sentences for the self-training. To ensure the quality of synthetic parallel data, we trained

[2] `BLEU+case.mixed+lang.[Task]+numrefs.1 +smooth.exp++test.wmt[Year]+tok.[Tok]+version.1.4.14`, Task=en-de/en-zh, Year=19/20, Tok=13a/zh

[3] `https://github.com/neulab/compare-mt`

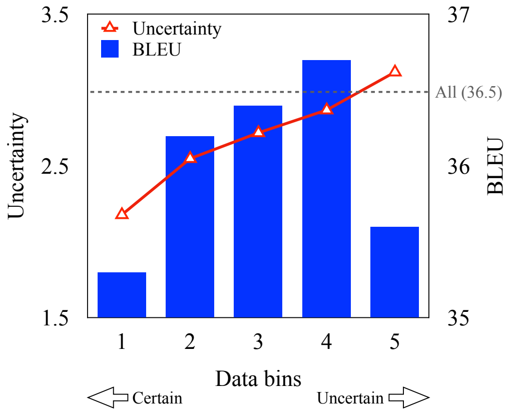


Figure 2: Relationship between uncertainty of monolingual data and the corresponding NMT performance. The BLEU score is averaged on WMT En⇒De newstest2019 and newstest2020.

a TRANSFORMER-BIG model for translating the source monolingual data to the target language. We generated translations using beam search with beam width 5, and followed Edunov et al. (2018)[4] to filter the generated sentence pairs (See Appendix A.1).

Self-training v.s. Data Size. We took a look at the performance of standard self-training and its relationship with data size. Figure 1 showed the results. Obviously, self-training with 8M synthetic data can already improve the NMT performance by a significant margin (36.2 averaged BLEU score on WMT En⇒De newstest2019 and newstest2020). Increasing the size of added monolingual data does not bring much more benefit. With all the 40M monolingual sentences, the final performance achieves only 36.5 BLEU points. It indicates that adding more monolingual data only is not a promising way to improve self-training, and more sophisticated approaches for exploiting the monolingual data are desired.

Self-training v.s. Uncertainty. In this experiment, we first adopted *fast-align*[5] to establish word alignments between source and target words in the authentic parallel corpus and used the alignments to build the bilingual dictionary $\mathcal{A}_b$. Then we used the bilingual dictionary to compute the data uncertainty expressed in Equation (3) for the sentences in the monolingual data set. After that, we ranked all the 40M monolingual sentences and grouped them into 5 equally-sized bins (i.e., 8M sentences per bin) according to their uncertainty scores. At last, we performed self-training with each bin of monolingual data.

We reported the translation performance in Figure 2. As seen, there is a trend of performance improvement with the increase of monolingual data uncertainty (e.g., bins 1 to 4) until the last bin. The last bin consists of sentences with excessively high uncertainty, which may contain erroneous synthetic target sentences. Training on these sentences forces the models to over-fit on these incorrect synthetic data, resulting in the confirmation bias issue (Arazo et al., 2020). These results corroborate with prior studies (Chang et al., 2017; Mukherjee and Awadallah, 2020) such that learning on certain examples brings little gain while on the excessively uncertain examples may also hurt the model training.

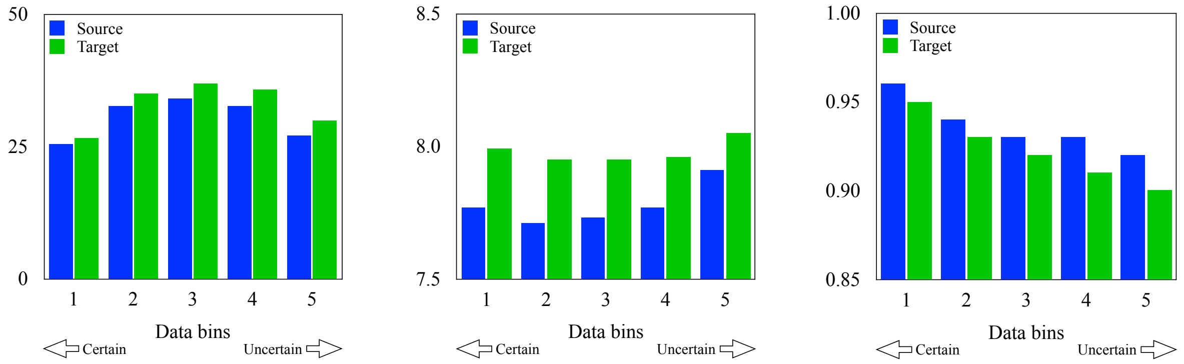
2.4 Linguistic Properties of Uncertain Data

We further analyzed the differences between the monolingual sentences with varied uncertainty to gain a deeper understanding of the uncertain data. Specifically, we performed linguistic analysis on the five data bins in terms of three properties: 1) sentence length that counts the tokens in the sentence, 2) word rarity (Platanios et al., 2019) that measures the frequency of words in a sentence with a higher value indicating a more rare sentence, and 3) translation coverage (Khadivi and Ney, 2005) that measures the ratio of source words being aligned with any target words. The first two reflect the properties of monolingual sentences while the last one reflects the quality of synthetic sentence pairs. We also presented the results of the synthetic target sentences for reference. Details of the linguistic properties are in Appendix A.2.

The results are reported in Figure 3. For the length property, we find that monolingual sentences with higher uncertainty are usually longer except for those with excessively high uncertainty (e.g., bin 5). The monolingual sentences in the last data bin noticeably contain more rare words than other bins in Figure 3(b), and the rare words in the sentences pose a great challenge in the NMT training process (Gu et al., 2020). In Figure 3(c), the overall coverage in bin 5 is the lowest among the self-training bins. In contrast, bin 1 with the lowest uncertainty has the highest coverage. These observations suggest that monolingual sentences in bin 1 indeed contain the easiest patterns while

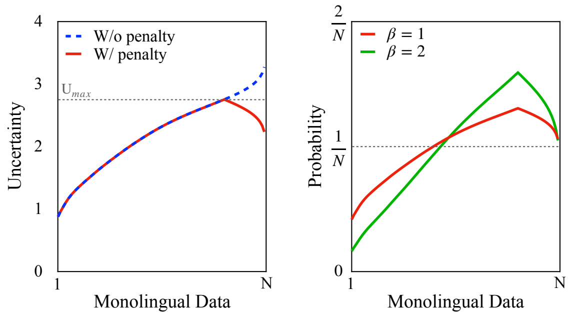
[4] https://github.com/pytorch/fairseq/tree/master/examples/backtranslation

[5] https://github.com/clab/fast_align



(a) Sentence Length (b) Word Rarity (c) Coverage

Figure 3: Comparison of monolingual sentences with varied uncertainty in terms of three properties, including sentence length, word rarity, and coverage.



(a) Uncertainty (b) Sampling Probability

Figure 4: Distribution of modified monolingual uncertainty and sampling probability. The sample with high uncertainty has more chance to be selected while that with excessively high uncertainty would be penalized.

monolingual sentences in bin 5 are the most difficult ones, which may explain their relatively weak performance in Figure 2.

3 Exploiting Monolingual Uncertainty

By analyzing the effect of monolingual data uncertainty on self-training in Section 2, we understood that monolingual sentences with relatively high uncertainty are more informative while also with high quality, which motivates us to emphasize the training on these sentences. In this section, we introduced the uncertainty-based sampling strategy for self-training and the overall framework.

3.1 Uncertainty-based Sampling Strategy

With the aforementioned measure of monolingual data uncertainty in Section 2.1, we propose the uncertainty-based sampling strategy for self-training, which prefers to sample monolingual sentences with relatively high uncertainty.

To ensure the data diversity and avoid the risk of being dominated by the excessively uncertain sentences, we sample monolingual sentences according to the uncertainty distribution with the highest uncertainty penalized. Specifically, given a budget of N_s sentences to sample, we set two hyperparameters to control the sampling probability as follows:

$$p = \frac{\left[\alpha \cdot \mathrm{U}(\mathbf{x}^j|\mathcal{A}_b)\right]^{\beta}}{\sum_{\mathbf{x}^j \in \mathcal{M}_x} \left[\alpha \cdot \mathrm{U}(\mathbf{x}^j|\mathcal{A}_b)\right]^{\beta}}, \tag{5}$$

$$\alpha = \begin{cases} 1, & \mathrm{U}(\mathbf{x}^j|\mathcal{A}_b) \leq \mathrm{U}_{max}, \\ max(\frac{2\mathrm{U}_{max}}{\mathrm{U}(\mathbf{x}^j|\mathcal{A}_b)} - 1, 0), & \text{else}, \end{cases} \tag{6}$$

where α is used to penalize excessively high uncertainty over a maximum uncertainty threshold U_{max} (See Figure 4(a)), the power rate β is used to adjust the distribution such that a larger β gives more probability mass to the sentences with high uncertainty (See Figure 4(b)).

The maximum uncertainty threshold U_{max} is assigned to the uncertainty value such that $R\%$ of sentences in the authentic parallel corpus have monolingual data uncertainty below than it. R is assumed to be as high as 80 to 100. Because for monolingual data with uncertainty higher than this threshold, they may not be translated correctly by the "teacher" model as there are inadequate such sentences in the authentic parallel data for the model to learn. As a result, monolingual sentences with uncertainty higher than U_{max} should be penalized in terms of the sampling probability.

Overall Framework. Figure 5 presents the framework of our uncertainty-based sampling for

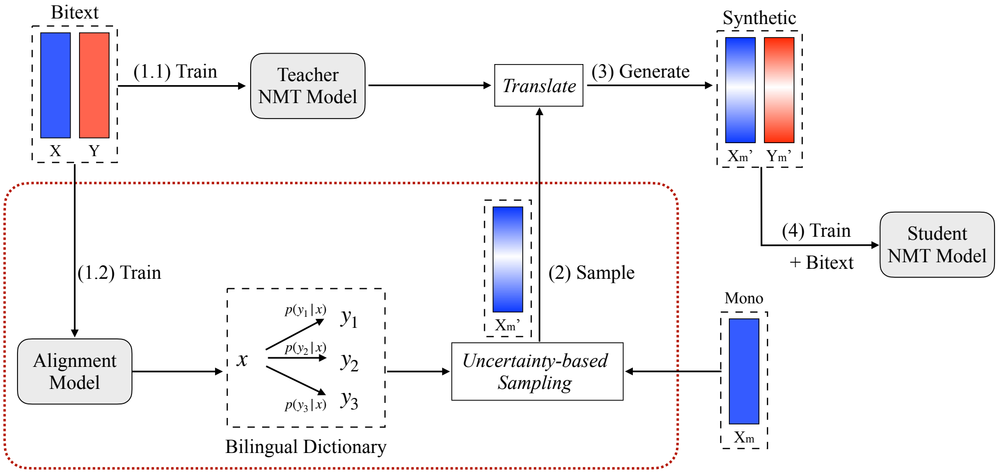


Figure 5: Framework of the proposed uncertainty-based sampling strategy for self-training. Procedures framed in the red dashed box corresponds to our approach integrated into the standard self-training framework. “Bitext”, “Mono”, “Synthetic” denotes authentic parallel data, monolingual data and synthetic parallel data, respectively.

self-training, which includes four steps: 1) train a “teacher” NMT model and an alignment model on the authentic parallel data simultaneously; 2) extract the bilingual dictionary from the alignment model and perform uncertainty-based sampling for monolingual sentences; 3) use the “teacher” NMT model to translate the sampled monolingual sentences to construct the synthetic parallel data; 4) train a “student” NMT model on the combination of synthetic and authentic parallel data.

3.2 Constrained Scenario

We first validated the proposed sampling approach in a constrained scenario, where we followed the experimental configuration in Section 2.3 with the TRANSFORMER-BASE model, the 8M bitext, and the 40M monolingual data. It allows the efficient evaluation of our approach with varied combinations of hyper-parameters and also the comparison with related methods. Specifically, we performed our approach by sampling 8M sentences from the 40M monolingual data and then combining the corresponding 8M synthetic data with the 8M bitext to train the TRANSFORMER-BASE model.

Table 1 reported the impact of β and R on the BLEU score. As shown, sampling with high uncertainty sentences and penalizing those with excessively high uncertainty improves translation performance from 36.6 to 36.9. In these experiments, the uncertainty threshold U_{max} for penalizing are 2.90 and 2.74, which are determined by the 90% and 80% (R=90 and 80 in Table 1) most certain sentences in the authentic parallel data, respectively.

BLEU		R		
		100	90	80
β	1	36.6	36.7	36.6
	2	36.7	**36.9**	36.6
	3	36.5	36.5	36.5

Table 1: Translation performance with respect to different values of β and R. The BLEU score is averaged on WMT En⇒De newstest2019 and newstest2020.

Obviously, the proposed uncertainty-based sampling strategy achieves the best performance with R at 90 and β at 2. In the following experiments, we use $R = 90$ and $\beta = 2$ as the default setting for our sampling strategy if not otherwise stated.

Effect of Sampling. Some researchers may doubt that the final translation quality is affected by the quality of the teacher model. Therefore, translations of high-uncertainty sentences should contain many errors, and it is better to add the results of oracle translations to discuss the sampling effect and the quality of pseudo-sentences separately. To dispel the doubt, we still used the aforementioned 8M bitext as the bilingual data, and used the rest of WMT19 En-De data (28.8M) as the held-out data (with oracle translations) for sampling. The results are listed in Table 2.

Clearly, our uncertainty-based sampling strategy (UNCSAMP) outperforms the random sampling strategy (RANDSAMP) when manual translations are used (Rows 2 vs. 3), demonstrating the effectiveness of our sampling strategy based on the un-

System	Data	En⇒De			En⇒Zh		
		2019	2020	Avg	2019	2020	Avg
Wu et al. (2019b)	Bitext	37.3	–	–	–	–	–
	+RandSamp	39.8	–	–	–	–	–
Shi et al. (2020)	Bitext	–	–	–	–	38.6	–
	+RandSamp	–	–	–	–	41.9	–
This Work	Bitext	39.6	31.0	35.3	37.1	42.5	39.8
	+RandSamp	41.6	33.1	37.3	37.6	43.8	40.7
	+SrcLM	41.7	33.1	37.4	37.3	44.0	40.7
	+UncSamp	$42.5^{\Uparrow}$	$34.4^{\Uparrow}$	**38.4**	$38.2^{\Uparrow}$	$44.3^{\uparrow}$	**41.3**

Table 4: Translation performance on WMT En⇒De and WMT En⇒Zh test sets. The results are reported with de-tokenized case-sensitive SacreBLEU. We adopt the Transformer-big with large batch training (Ott et al., 2018) to achieve the strong performance. "↑ / ⇑": indicate statistically significant improvement over RandSamp $p < 0.05/0.01$ respectively.

#	Data	2019	2020	Avg
1	Bitext	36.9	27.7	32.3
2	+ RandSamp Ora	37.4	28.0	32.7
3	+ UncSamp Ora	37.8	28.2	33.0
4	+ RandSamp ST	40.0	30.1	35.0
5	+ UncSamp ST	40.4	30.5	**35.4**

Table 2: Comparison of our UncSamp and RandSamp with manual translations (Ora: manual translations; ST: pseudo-sentences) on WMT En⇒De newstest2019 and newstest2020.

Data	2019	2020	Avg
RandSamp	40.9	31.6	36.2
DWF	39.6	30.1	34.8
SrcLM	41.1	32.0	36.5
UncSamp	41.6	32.3	36.9
+ Filtering	41.5	32.7	**37.1**

Table 3: Comparison of the proposed uncertainty-based sampling strategy with related methods on WMT En⇒De newstest2019 and newstest2020.

certainty. Another interesting finding is that using the pseudo-sentences outperforms using the manual translations (Rows 4 vs. 2, 5 vs. 3). One possible reason is that the Transformer-big model to construct the pseudo-sentences was trained on the whole WMT19 En-De data that contains the held-out data, which serves as self-training to decently improve the supervised baseline (He et al., 2019).

Comparison with Related Work. We compared our sampling approach with two related works, i.e., difficult word by frequency (DWF, Fadaee and Monz, 2018) and source language model (SrcLM, Lewis, 2010). The former one was proposed for monolingual data selection for back-translation, in which sentences with low-frequency words were selected to boost the performance of back-translation. The latter one was proposed for in-domain data selection for in-domain language models. Details of the implementation of related work are in Appendix A.3.

Table 3 listed the results. For DWF, it brings no improvement over RandSamp, indicating that the technique developed for back-translation may not work for self-training. As for SrcLM, it achieves a marginal improvement over RandSamp. The proposed UncSamp approach outperforms the baseline RandSamp by +0.7 BLEU point, which demonstrates the effectiveness of our approach. In addition to our UncSamp approach, we also utilized another n-gram language model at the target side to further filter out the synthetic data with potentially erroneous target sentences. By filtering out 20% sentences from the sampled 8M sentences, our UncSamp approach achieves a further improvement up to +0.9 BLEU point.

3.3 Unconstrained Scenario

We extended our sampling approach to the unconstrained scenario, where the scale of data and the capacity of NMT models for self-training are increased significantly. We conducted experiments on the high-resource En⇒De and En⇒Zh translation tasks with all the authentic parallel data, including 36.8M sentence pairs for En⇒De and 22.1M for En⇒Zh, respectively. For monolingual data, we considered all the 200M English newscrawl monolingual data to perform sampling. We trained the TRANSFORMER-BIG model for experiments.

Table 4 listed the main results of large-scale self-training on high-resource language pairs. As shown, our TRANSFORMER-BIG models trained on the authentic parallel data achieve the performance competitive with or even better than the submissions to WMT competitions. Based on such strong baselines, self-training with RANDSAMP improves the performance by +2.0 and +0.9 BLEU points on En⇒De and En⇒Zh tasks respectively, demonstrating the effectiveness of the large-scale self-training for NMT models. With our uncertainty-based sampling strategy UNCSAMP, self-training achieves further significant improvement by +1.1 and +0.6 BLEU points over the random sampling strategy, which demonstrates the effectiveness of exploiting uncertain monolingual sentences.

Unc	**BITEXT**	**RANDSAMP**	**UNCSAMP**	
			BLEU	△(%)
Low	38.1	39.7	41.5	8.9
Med	34.2	36.7	37.4	9.3
High	31.0	33.4	34.4	**10.9**

Table 5: Translation performance on uncertain sentences. The relative improvements over BITEXT for UNCSAMP are also presented.

Freq	**BITEXT**	**RANDSAMP**	**UNCSAMP**	
			Fmeas	△(%)
Low	52.3	53.8	54.7	**4.5**
Med	65.2	66.5	66.9	2.6
High	70.3	71.6	72.0	2.4

Table 6: Prediction accuracy of low-frequency words in the translation outputs. The relative improvements over BITEXT for UNCSAMP are also presented.

3.4 Analysis

In this section, we conducted analyses to understand how the proposed uncertainty-based sampling approach improved the translation performance. Concretely, we analyzed the translation outputs of WMT En⇒De newstest2019 from the TRANSFORMER-BIG model in Table 4.

Uncertain Sentences. As we propose to enhance high uncertainty sentences in self-training, one remaining question is whether our UNCSAMP approach improves the translation quality of high uncertainty sentences. Specifically, we ranked the source sentences in the newstest2019 by the monolingual uncertainty, and divided them into three equally sized groups, namely Low, Medium and High uncertainty.

The translation performance on these three groups is reported in Table 5. The first observation is that sentences with high uncertainty are with relatively low BLEU scores (i.e., 31.0), indicating the higher difficulty for NMT models to correctly decode the source sentences with higher uncertainty. Our UNCSAMP approach improves the translation performance on all sentences, especially on the sentences with high uncertainty (+10.9%), which confirms our motivation of emphasizing the learning on uncertain sentences for self-training.

Low-Frequency Words. Partially motivated by Fadaee and Monz (2018), we hypothesized that the addition of monolingual data in self-training has the potential to improve the prediction of low-frequency words at the target side for the NMT models. Therefore, we investigated whether our approach has a further boost to the performance on the prediction of low-frequency words. We calculated the word accuracy of the translation outputs with respect to the reference in newstest2019 by `compare-mt`. Following Wang et al. (2020), we divided words into three categories based on their frequency, including High: the most 3,000 frequent words; Medium: the most 3,001-12,000 frequent words; Low: the other words.

Table 6 listed the results of word accuracy on these three groups evaluated by F-measure. First, we observe that low-frequency words in BITEXT are more difficult to predict than medium- and high-frequency words (i.e., 52.3 v.s. 65.2 and 70.3), which is consistent with Fadaee and Monz (2018). Second, adding monolingual data by self-training improves the prediction performance of low-frequency words. Our UNCSAMP approach outperforms RANDSAMP significantly on the low-frequency words. These results suggest that emphasizing the learning on uncertain monolingual sentences also brings additional benefits for the learning of low-frequency words at the target side.

4 Related Work

Synthetic Parallel Data. Data augmentation by synthetic parallel data has been the most simple and effective way to utilize monolingual data for NMT, which can be achieved by self-training (He et al., 2019) and back-translation (Sennrich et al., 2016a). While back-translation has dominated the NMT area for years (Fadaee and Monz, 2018; Edunov et al., 2018; Caswell et al., 2019), recent works on translationese (Marie et al., 2020; Graham et al., 2019) suggest that NMT models trained with back-translation may lead to distortions in automatic and human evaluation. To address the problem, starting from WMT2019 (Barrault et al., 2019), the test sets only include naturally occurring text at the source-side, which is a more realistic scenario for practical translation usage. In this new testing setup, the forward-translation (Zhang and Zong, 2016), i.e., self-training in NMT, becomes a more promising method as it also introduces naturally occurring text at the source-side. Therefore, we focus on the data sampling strategy in the self-training scenario, which is different from these prior studies.

Data Uncertainty in NMT. Data uncertainty in NMT has been investigated in the last few years. Ott et al. (2018) analyzed the NMT models with data uncertainty by observing the effectiveness of data uncertainty on the model fitting and beam search. Wang et al. (2019) and Zhou et al. (2020) computed the data uncertainty on the back-translation data and the authentic parallel data and proposed uncertainty-aware training strategies to improve the model performance, respectively. Wei et al. (2020) proposed the uncertainty-aware semantic augmentation method to bridge the discrepancy of the data distribution between the training and the inference phases. In this work, we propose to explore monolingual data uncertainty to perform data sampling for the self-training in NMT.

5 Conclusion

In this work, we demonstrate the necessity of distinguishing monolingual sentences for self-training in NMT, and propose an uncertainty-based sampling strategy to sample monolingual data. By sampling monolingual data with relatively high uncertainty, our method outperforms random sampling significantly on the large-scale WMT English⇒German and English⇒Chinese datasets. Further analyses demonstrate that our uncertainty-based sampling approach does improve the translation quality of high uncertainty sentences and also benefits the prediction of low-frequency words at the target side. The proposed technology has been applied to TranSmart[6] (Huang et al., 2021), an interactive machine translation system in Tencent, to improve the performance of its core translation engine. Future work includes the investigation on the confirmation bias issue of self-training and the effect of decoding strategies on self-training sampling.

Acknowledgments

This work is partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 2410021, Research Impact Fund (RIF), R5034-18; CUHK 14210717, General Research Fund), and Tencent AI Lab RhinoBird Focused Research Program (GF202036). We sincerely thank the anonymous reviewers for their insightful suggestions on various aspects of this work.

References

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation. In *WMT*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. In *ACL*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *WMT*.

Haw-Shiuan Chang, Erik G Learned-Miller, and Andrew McCallum. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *NeurIPS*.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

[6] https://transmart.qq.com/index

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *EMNLP*.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *EMNLP*.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *arXiv*.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *EMNLP*.

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

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. In *ICLR*.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *EMNLP*.

Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: a practical interactive machine translation system. *arXiv*.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020a. Exploiting unsupervised data for emotion recognition in conversations. In *EMNLP: Findings*.

Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael R Lyu, and Zhaopeng Tu. 2020b. Data rejuvenation: Exploiting inactive training examples for neural machine translation. In *EMNLP*.

Shahram Khadivi and Hermann Ney. 2005. Automatic filtering of bilingual corpora for statistical machine translation. In *NLDB*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *EMNLP*.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *EMNLP*.

Robert C Moore William Lewis. 2010. Intelligent selection of language model training data. *ACL*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv*.

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. *TACL*.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *ACL*.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. In *NeurIPS*.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *NAACL (Demonstrations)*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. In *WMT*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *ICML*.

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 *NAACL (Demonstrations)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *NAACL*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *WMT 2018*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *ACL*.

Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, and Jie Hao. 2020. Oppo's machine translation systems for wmt20. In *WMT*.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *CVPR*.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In *ACL*.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Xu Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *ACL*.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *ACL*.

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 *NeurIPS*.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. In *EMNLP*.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the Inference Calibration of Neural Machine Translation. In *ACL*.

Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Luxi Xing, and Weihua Luo. 2020. Uncertainty-aware semantic augmentation for neural machine translation. In *EMNLP*.

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. 2019a. Pay less attention with lightweight and dynamic convolutions. In *ICLR*.

Lijun Wu, Yiren Wang, Yingce Xia, QIN Tao, Jianhuang Lai, and Tie-Yan Liu. 2019b. Exploiting monolingual data at scale for neural machine translation. In *EMNLP*.

Shuangzhi Wu, Xing Wang, Longyue Wang, Fangxu Liu, Jun Xie, Zhaopeng Tu, Shuming Shi, and Mu Li. 2020. Tencent neural machine translation systems for the wmt20 news translation task. In *WMT*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *AAAI*.

Chunting Zhou, Graham Neubig, and Jiatao Gu. 2019. Understanding knowledge distillation in non-autoregressive machine translation. In *ICLR*.

Yikai Zhou, Baosong Yang, Derek F Wong, Yu Wan, and Lidia S Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *ACL*.

A Appendix

A.1 Synthetic Data

When performing self-training, we constructed the synthetic data by translating the monolingual sentences via beam search with beam width 5, and followed Edunov et al. (2018)[7] to remove sentences longer than 250 words as well as sentence-pairs with a source/target length ratio exceeding 1.5. The "teacher" NMT model for self-training is the TRANSFORMER-BIG model to ensure the quality of synthetic data.

A.2 Linguistic Properties

Word Rarity. Word rarity measures the frequency of words in a sentence with a higher value indicating a more rare sentence (Platanios et al., 2019). The word rarity of a sentence is calculated as follows:

$$\mathrm{WR}(\mathbf{x}) = -\frac{1}{T_x}\sum_{t=1}^{T_x}\log p(x_t), \tag{7}$$

where $p(x_t)$ denotes the normalized frequency of word x_t in the authentic parallel data, and T_x is the sentence length.

Coverage. Coverage measures the ratio of source words being aligned by any target words (Tu et al., 2016). Firstly, we trained an alignment model on the authentic parallel data by *fast-align*[8]. Then we used the alignment model to force-align the monolingual sentences and the synthetic target sentences. Next, we calculated the coverage of each source sentence, and report the averaged coverage of each data bin. The lower coverage of monolingual sentences in bin 5 indicates that they are not aligned as well as the other bins.

A.3 Comparison with Related Work

We compared our sampling approach with two related works, i.e., difficult word by frequency (DWF, Fadaee and Monz, 2018) and source language model (SRCLM, Lewis, 2010). The former one was proposed for monolingual data selection for back-translation, in which sentences with low-frequency words were selected to boost the performance of back-translation. The latter one was proposed for in-domain data selection for in-domain language models.

For DWF, we ranked the monolingual data by word rarity (Platanios et al., 2019) of sentences and also selected the top 80M monolingual data for self-training. For SRCLM, we trained an N-gram language model (Heafield, 2011)[9] on the source sentences in the bitext and measured the distance between each monolingual sentence to the bitext source sentences by cross-entropy. Similarly, we selected 8M monolingual data with the lowest cross-entropy for self-training.

[7] https://github.com/pytorch/fairseq/tree/master/examples/backtranslation

[8] https://github.com/clab/fast_align

[9] https://kheafield.com/code/kenlm/