Improving Simultaneous Machine Translation with Monolingual Data

Hexuan Deng[1]*, Liang Ding[2], Xuebo Liu[1]†, Meishan Zhang[1], Dacheng Tao[2], Min Zhang[1]

[1] Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
[2] JD Explore Academy, JD.com Inc.
22s051030@stu.hit.edu.cn, dingliang1@jd.com, liuxuebo@hit.edu.cn,
zhangmeishan@hit.edu.cn, dacheng.tao@gmail.com, zhangmin2021@hit.edu.cn

Abstract

Simultaneous machine translation (SiMT) is usually done via sequence-level knowledge distillation (Seq-KD) from a full-sentence neural machine translation (NMT) model. However, there is still a significant performance gap between NMT and SiMT. In this work, we propose to leverage monolingual data to improve SiMT, which trains a SiMT student on the combination of bilingual data and external monolingual data distilled by Seq-KD. Preliminary experiments on En⇒Zh and En⇒Ja news domain corpora demonstrate that monolingual data can significantly improve translation quality (e.g., +3.15 BLEU on En⇒Zh). Inspired by the behavior of human simultaneous interpreters, we propose a novel monolingual sampling strategy for SiMT, considering both chunk length and monotonicity. Experimental results show that our sampling strategy consistently outperforms the random sampling strategy (and other conventional typical NMT monolingual sampling strategies) by avoiding the key problem of SiMT – hallucination, and has better scalability. We achieve +0.72 BLEU improvements on average against random sampling on En⇒Zh and En⇒Ja. Data and codes can be found at https://github.com/hexuandeng/Mono4SiMT.

Introduction

Simultaneous machine translation (SiMT) (Gu et al. 2017; Ma et al. 2019; Arivazhagan et al. 2019; Zheng et al. 2020) has been proposed to generate real-time translation by starting decoding before the source sentence ends. However, generation conditioned on the partial source sentence prevents a model from properly capturing the whole semantics, especially for distant languages, e.g., English and Japanese (He et al. 2015; Chen et al. 2021). In response to this problem, motivated by the recent success of non-autoregressive translation, sequence-level knowledge distillation (Seq-KD, Kim and Rush 2016) becomes the preliminary step for training SiMT models, with a full-sentence neural machine translation (NMT) model as the teacher (Ren et al. 2020; Zhang, Feng, and Li 2021), which helps to generate monotonous knowledge by reducing data complexity (Zhou, Gu, and Neubig 2020).

Although Seq-KD narrows the gap between full-sentence NMT teachers and SiMT students, the performance gap is still significant. Techniques like self-training (Zhang and Zong 2016; Jiao et al. 2021) are known to effectively improve machine translation performance by using large-scale monolingual data. However, to the best of our knowledge, improving SiMT through semi-supervised learning has not been well validated yet.

To this aim, we leverage the monolingual data to perform Seq-KD and train the SiMT student model on the combination of distilled monolingual and bilingual data. Exploiting monolingual data for SiMT provides appealing benefits. First, the monolingual data and bilingual data in machine translation are generally complementary to each other (Sennrich, Haddow, and Birch 2016a; Zhang and Zong 2016; Zhou and Keung 2020; Ding et al. 2022). Accordingly, using monolingual for SiMT transfers both the knowledge of the bilingual data (implicitly encoded in the full-sentence NMT teacher) and that of monolingual data, maintaining the merit of Seq-KD to reduce the complexity of the bilingual data. Secondly, the amount of available monolingual data is several orders of magnitude larger than that of bilingual data, offering great potential to enjoy attractive expandability.

However, unlike NMT, it is difficult for SiMT to handle long-distance reordering (Zhou and Keung 2020). Therefore, the pseudo-targets generated by the full-sentence NMT teacher model are not always suitable for SiMT. Inspired by strategies used in human simultaneous interpretation, e.g., finer segments and monotonic alignments (He, Boyd-Graber, and Daumé III 2016), we propose novel strategies for sampling monolingual data suitable for SiMT, considering both the chunk lengths and monotonicity. We validate our strategy on several large-scale datasets of news domain (En⇒Zh and En⇒Ja). Our contributions are as follows:

- We empirically demonstrate that using monolingual data is beneficial to SiMT systems.
- Our monolingual data sampling strategy for SiMT significantly outperforms the random sampling and conventional NMT monolingual sampling strategies, especially evaluating at low latency.
- Our strategy effectively alleviates the key issue of SiMT, i.e., hallucination problem, and has high expandability, e.g., enlarging the scale of monolingual data consistently improves performance.

*Work was done when Hexuan was interning at JD Explore Academy.

†Corresponding author: Xuebo Liu.

The paper is an early step in exploring monolingual data for SiMT, which can narrow the performance gap between SiMT models and the SOTA full-sentence NMT models. We hope the promising effect of the monolingual sampling strategy on SiMT can encourage further investigation and pave the way toward more effective SiMT models.

Background and Related Work

Simultaneous Machine Translation Full-sentence NMT models use Seq2seq framework, where the encoder takes the source sentence $\boldsymbol{x} = (x_1, ..., x_m)$ as input, and outputs hidden state $\boldsymbol{h} = (h_1, ..., h_m)$. Then, the decoder iteratively predicts the next token y_t based on the hidden state and previously generated tokens until the end of the sequence:

$$\hat{y_t} = \underset{y_t}{\operatorname{argmax}}\, p\left(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}; \boldsymbol{\theta}\right) \tag{1}$$

In SiMT, we cannot access the entire source sentence when decoding. Ma et al. (2019) propose a simple but efficient wait-k policy to balance translation quality and delay. Specifically, it first reads k words, then loops to read and write a word until the end of the sentence:

$$\hat{y_t} = \underset{y_t}{\operatorname{argmax}}\, p\left(y_t \mid \boldsymbol{x}_{\leq g_{\text{wait-}k}(t)}, \boldsymbol{y}_{<t}; \boldsymbol{\theta}\right) \tag{2}$$

where $g_{\text{wait-}k}(t) = \min\{k + t - 1, |\boldsymbol{x}|\}$ indicates the number of source words that can be seen when predicting word y_t under the wait-k policy.

Several works have been proposed to narrow the gap between SiMT and NMT datasets. He et al. (2015) use handwriting language-specific rules based on syntax trees to generate pseudo-targets with fewer reordering, but it requires linguistic knowledge and is difficult to transfer to other language pairs. Zhang et al. (2020) use the sentence-aligned parallel corpus to train an NMT model and generate pseudo-targets with a policy according to the attention of NMT, while Chen et al. (2021) directly use the test-time wait-k policy, which significantly reduces the anticipation rate while simplifies the computational complexity. Han et al. (2021) employ a method based on chunk-wise reordering and NAT refinement to generate monotonic and smooth references. Unlike the above approaches that utilize bilingual data effectively, our study is the first work to investigate how to improve SiMT with large-scale monolingual data, which is orthogonal to the above approaches.

Semi-Supervised NMT NMT models are data hungry, and the translation quality highly depends on the quality and quantity of parallel corpus (Koehn and Knowles 2017; Liu et al. 2020a). Researchers thus turn to investigate the effects of using large-scale monolingual data (Zhang and Zong 2016; Domhan and Hieber 2017; Edunov et al. 2018; Ding and Tao 2021) with semi-supervised learning (Zhu and Goldberg 2009). The general process follows several steps: 1) train a base model with bilingual data; 2) decode the large-scale monolingual data with the pre-trained base model to obtain the synthetic data; and 3) retrain the model with the concatenation of bilingual and synthetic data. Designing an effective monolingual sampling strategy is at the core of the process. Moore and Lewis (2010) select in-domain monolingual samples through the source language model. Fadaee and Monz (2018) improve the prediction accuracy of the model by selecting sentences with lower frequency words, while Jiao et al. (2021) achieve a similar purpose by sampling monolingual data with high uncertainty. While semi-supervised learning shows great success in full-sentence translation, few works explore the effects of using monolingual data for SiMT. We take the first step to investigate SiMT-aware monolingual sampling strategies and their best combination and provide a comprehensive discussion to show the scalability of our approach.

Monolingual Data Sampling Strategies

We introduce the sampling strategies and the corresponding metrics, where monolingual data with lower scores are considered more efficient and used for training. The tendency to choose longer sentences is added to all these strategies and will be introduced first.

Sentence Length Longer sentences usually contain more information, encouraging the model to make use of more context information (Platanios et al. 2019). Besides, training with longer sentences can suppress the generation of end signal "<EOS>" and nicely alleviate the early-stop phenomenon in SiMT, where the generating ends are given the incomplete source input. Therefore, in all subsequent sampling strategies, we add long sentence tendency factor α by replacing the sentence length term (or similar item) $|\boldsymbol{x}|$ with $|\boldsymbol{x}|^{\alpha}$ (or $|\boldsymbol{x}|^{1/\alpha}$), aiming at tending to choose longer sentences while maintaining the effectiveness of the strategy. In our experiments, we set $\alpha = 0.5$ as default.

Sample Corpora More Suitable for SiMT

In response to different word order between language pairs, He, Boyd-Graber, and Daumé III (2016) point out that human interpretation often: 1) breaks source sentences into multiple smaller chunks and uses conjunctions for fluently connecting; 2) uses passivization to wait for the source to give the verb without stopping the translation process, especially when from head-final languages (e.g., Japanese) to head-initial languages (e.g., English). Both of them greatly alleviate the problems above while ensuring fluency.

Chunk Length-Based Strategy Inspired by the first phenomenon, the easiest way is to select data with shorter chunks for training to develop its tendencies, aiming at obtaining the same benefits as above. As for chunk extraction, we want to evaluate the chunk length of the current monolingual corpora at the lowest cost rather than extracting meaningful units. Under such consideration, we propose the following two metrics to give a relatively accurate evaluation.

Inspired by Chiang (2007), *Alignment-based approach* selects the shortest contiguously aligned block as a chunk, which satisfies that tokens in the source part are aligned with and only with corresponding tokens in the target part and vice versa, while the source part and the target part are contiguous and inseparable. As shown in Figure 1, the parts enclosed by the red box are chunks we identified. This

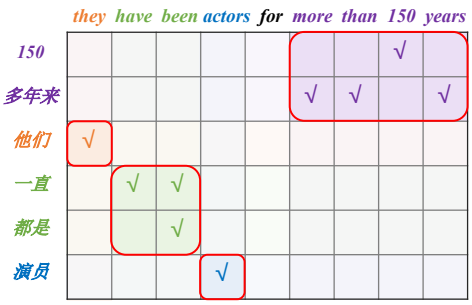

Figure 1: Example of alignment-based chunk extraction, where "✓" means that the source- and target-side tokens are aligned, and the red rectangles are the extracted chunk pairs.

method can extract meaningful chunks in most cases but need pseudo-targets and alignments for monolingual data, which is time-consuming.

To extract chunks efficiently, inspired by Sen, Germann, and Haddow (2021), we employ source-side language model (LM): *LM-based approach* keeps track of the LM score of the prefix of source sentences and adds token once at a time. If the new LM score is lower than the previous one, the previous prefix will be considered as a chunk. Afterwards, the next word is regarded as the beginning of the sentence, and recursively perform the above steps until the end of the sentence. Although there is no information about pseudo-targets, it can also play a similar or even better effect than the previous method in our experiments (See Table 2).

In the calculation of metrics, the numerator is the number of alignments in the source sentence for the alignment-based approach and sentence length for the LM-based approach. We add index α to those numerators as exponents to reflect the long sentence tendency. In this way, for the alignment-based approach, sentences with denser alignments are also tended to be chosen, which intuitively have lower error rates and contain more information, which should also be encouraged. Formally, if we define the total number of chunks in the sentence as c and the numerator as ℓ, the chunk length-based metric for the sentence is:

$$S_{chunk} = \frac{\ell^{\alpha}}{c} \tag{3}$$

Monotonicity-Based Strategy Inspired by the second phenomenon, we take a straightforward solution to choose sentences with more monotonous alignments directly. Refer to Chen et al. (2021), we use k-Anticipation Rate (k-AR) as metric for monotonicity. Specifically, for each aligned target word y_j, it is considered a k-anticipation if it is aligned to a source word x_i that is k words behind. The k-AR is then calculated as the percentage of k-anticipation among all aligned word pairs. Specifically, if the set $\mathcal{A} = \{(i_t, j_t)\}_{t=1}^{N}$ represents all aligned token-pairs $\boldsymbol{x}_{i_k} \sim \boldsymbol{y}_{j_k}$, the monotonicity-based metric for the sentence is:

$$S_{mono} = \frac{1}{|\mathcal{A}|^{1/\alpha}} \sum_{t=1}^{|\mathcal{A}|} \mathbb{1}[i_t \leq j_t + k] \tag{4}$$

where α is the long sentence tendency factor, which also adds bias for sentences with denser alignments as with the alignment-based approach.

Sentence Difficulty

In traditional NMT, there are some solutions for sampling monolingual data according to difficulty. We choose two of them and add the same long sentence tendency factor α for comparison.

Fadaee and Monz (2018) propose that monolingual data containing low-frequency words are more conducive to model training. Then Platanios et al. (2019) use the source-side unigram language model to reflect the tendency to select sentences that are longer and contain more low-frequency words at the same time. In our setup, for monolingual sentence $\boldsymbol{x} = (x_1, ..., x_m)$, and the probability $\hat{p}(x_i)$ of each word x_i occurred in the bilingual corpora, taking into account the tendency to choose long sentences, the frequency metric for the sentence is:

$$S_{\text{rarity}} = -\frac{1}{|\boldsymbol{x}|^{\alpha}} \sum_{i=1}^{|\boldsymbol{x}|} \log \hat{p}(x_i) \tag{5}$$

Jiao et al. (2021) propose a metric based on uncertainty. It first evaluates word level entropy E by using the alignment $\mathcal{A}$ on bilingual corpora to capture the translation modalities of each source token. Specifically, for a given monolingual sentence $\boldsymbol{x} = (x_1, ..., x_m)$, if $\mathcal{A}(x_i)$ records all possible target tokens y_j aligned with source token x_i, and calculate the translation probability $p(y_j|x_i)$ according to it, the word level entropy is:

$$E(y \mid \mathcal{A}, x_i) = -\sum_{y_j \in \mathcal{A}(x_i)} p(y_j \mid x_i) \log p(y_j \mid x_i)$$

For the monolingual data, taking into account the tendency to choose long sentences, its uncertainty metric is:

$$S_{\text{uncer}} = \frac{1}{|\boldsymbol{x}|^{\alpha}} \sum_{i=1}^{|\boldsymbol{x}|} E(y \mid \mathcal{A}, x = x_i) \tag{6}$$

Experiments

Experimental Setup

Bilingual Data We conduct experiments on two widely-used SiMT language directions: English-Chinese (En⇒Zh) and English-Japanese (En⇒Ja). To make the experiments convincing, we select resource-rich datasets of news domain: For En⇒Zh, we use CWMT Corpus[1] (Chen and Zhang 2019) as training data, NJU-newsdev2018 as the validation set and report results on CWMT2008, CWMT2009, and CWMT2011; For En⇒Ja, we use JParaCrawl[2] (Morishita, Suzuki, and Nagata 2020) and WikiMatrix[3] (Schwenk et al. 2021) as training data, newsdev2020 as the validation

[1] http://nlp.nju.edu.cn/cwmt-wmt/
[2] https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/
[3] https://opus.nlpl.eu/WikiMatrix.php

	Raw	KD	KD+Mono.
	Teacher: 48.55		
wait-1	28.62	29.93	**35.64**
wait-3	35.39	36.15	**39.82**
wait-5	39.07	41.14	**43.46**
wait-7	42.52	43.76	**45.95**
wait-9	44.02	45.66	**47.51**
Avg.	37.92 (- / -)	39.33 **(+1.41/-)**	**42.48 (+4.56/+3.15)**

Table 1: The effects of using monolingual data. "Raw/ KD" means the results of original/distilled parallel data, and "+Mono." represents enhancing the model with synthetic data generated by randomly sampled monolingual data. Gains against "Raw" and "KD" are given separately below the underline. Average scores on all delays are underlined. The best results are bold.

set and report results on newstest2020. Considering the corpora are noisy, we apply a series of filtration rules to them, including 1) empty and duplicated lines, 2) sentence pairs with invalid characters, 3) sentence pairs with too many or too few words, and 4) those with too large bilingual length ratios, etc. After data cleaning, we randomly select a subset of $7M$ sentence pairs as training data for both En⇒Zh and En⇒Ja. We use SentencePiece (Kudo and Richardson 2018) to split the training data into subword units (Sennrich, Haddow, and Birch 2016b) with 32K merge operations. We publicly release our processed datasets[4].

Monolingual Data We closely follow previous works to randomly select monolingual data from publicly available News Crawl corpus[5] (Zhang and Zong 2016; Wu et al. 2019). For a fair comparison, the monolingual data used in the main experiments have the same size as the corresponding bilingual data, i.e., $7M$. To comprehensively investigate the effects of different monolingual sampling strategies in Table 2, we randomly sample up to $42M$ English data from News Crawl 2016 and 2017 in the main experiments. For the at-scale experiments in Table 5, we randomly sample up to $540M$ sentences from News Crawl 2007∼2017 and News Discussions 2014∼2017.

Model Training We closely follow previous SiMT works (Ren et al. 2020; Zhang, Feng, and Li 2021; Fukuda et al. 2021; Liu et al. 2021a; Zhao et al. 2021) to adopt sequence-level knowledge distillation (Kim and Rush 2016) for all systems. Specifically, we train a full-sentence BASE Transformer (Vaswani et al. 2017) as the teacher on the original bilingual dataset, then perform beam-search decoding for the source side of the original bilingual data or newly introduced monolingual data to generate the distilled data. The student SiMT model follows the BASE model, except for using causal encoders and wait-k policy. To investigate the effects of a better teacher, we use full-sentence BIG Transformer at Table 6. Note that we train all models with identical training steps.

[4]https://drive.google.com/drive/folders/1HbzxBD0klgX-EugVGB36CFVdObJJ5Uk7?usp=sharing

[5]http://data.statmt.org/news-crawl

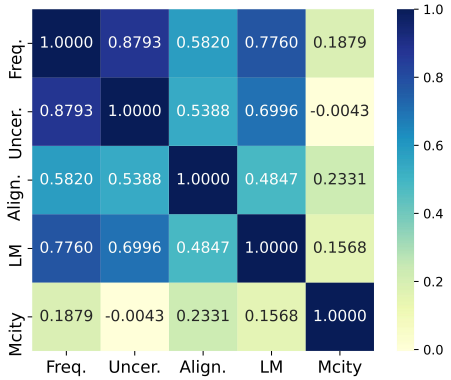

Figure 2: Covariance matrix between different sampling strategies. We score the monolingual dataset using different strategy metrics and calculate the correlation coefficient between different scores. "Freq." and "Uncer." are sentence difficulty metrics estimated with word frequency and uncertainty, respectively. "Align." and "LM" are chunk length metrics using alignment-based and LM-based approaches, respectively. "Mcity" is monotonicity metric using 3-anticipation rate. The same notations are used in following-up tables.

We use the *SacreBLEU* (Post 2018) to measure the translation quality and *SimulEval* (Ma et al. 2020) to measure the latency for each delay under the *wait-k* (Ma et al. 2019) policy, and also report the averaged BLEU for different delays to avoid stochasticity. The CWMT test sets have up to 3 references. Thus, we report the 3-reference BLEU score. We use *fast-align* (Dyer, Chahuneau, and Smith 2013) to extract the alignment information for sentences in Table 4, and strategies more suitable for SiMT, and use *KenLM* (Heafield et al. 2013) to calculate source language model score in chunk length-based strategy.

Empirical Findings

In this section, we comprehensively conduct preliminary studies on CWMT En⇒Zh to show 1) the necessity of using monolingual data, 2) the superiority of our proposed SiMT-aware monolingual sampling strategies, and 3) the best strategy combination as our default method.

Monolingual data significantly improves SiMT. In order to explore the effect of adding monolingual data, we add the synthetic data generated by randomly sampled monolingual sentences to the distilled parallel data with a ratio of 1:1. We report the results of original parallel data ("Raw") for reference. As shown in Table 1, we can see that distillation improves the SiMT with +1.41 BLEU points on average, and leveraging the randomly sampled monolingual data further pushes the BLEU points by a large margin, i.e., +3.15, especially for the low-latency settings, e.g., +5.71 for wait-1. This confirms the effectiveness of monolingual data for SiMT and urges us to investigate better sampling strategies for monolingual data.

Strategy	*wait-1*	*wait-3*	*wait-5*	*wait-7*	*wait-9*	*Avg.*	Δ
Random	35.64	39.82	43.46	45.95	47.51	42.48	
Frequency-Based Sentence Difficulty Strategy	**36.69**	40.78	44.11	46.12	47.76	43.09	+0.61
Uncertainty-Based Sentence Difficulty Strategy	36.26	40.95	43.33	46.30	47.57	42.88	+0.40
Alignment-Based Chunk Length Strategy	36.62	41.20	43.68	**46.85**	**48.05**	**43.28**	**+0.80**
LM-Based Chunk Length Strategy	36.37	**41.70**	**44.12**	45.92	47.94	43.21	+0.73
Monotonicity-Based Strategy	35.97	40.25	42.88	45.65	46.80	42.31	-0.17

Table 2: The effect of different sampling strategies. Since our proposed strategy and baseline belong to the same policy, there is almost no difference in latency. Therefore, we display the results in the form of table to highlight the details of the improvement in translation quality. Improvements against random sampling "Random" are in column **Δ**.

	Chunk (Align.)	**+Mcity**	**Chunk (LM)**	**+Mcity**
wait-1	36.62 (-)	**37.08 (+0.46)**	36.37 (-)	**37.40 (+1.03)**
wait-3	**41.20**	41.10	**41.70**	40.49
wait-5	43.68	**44.28**	44.12	**44.44**
wait-7	**46.85**	46.46	45.92	**46.27**
wait-9	**48.05**	47.69	47.94	**48.00**
Avg.	43.28 (-)	**43.32 (+0.04)**	43.21 (-)	**43.32 (+0.11)**

Table 3: The complementary effect of chunk length-based strategies, i.e., "Chunk (Align.)" and "Chunk (LM)", and monotonicity-based strategy "+Mcity". We combine the strategies with significant differences (Covariance$<$0.3) according to correlation analysis in Figure 2: "+Mcity" with alignment based chunk length strategy "Align." and language model based chunk length strategy "LM".

SiMT-aware sampling strategies do help. We test the effects of our deliberately designed strategies for SiMT. As shown in Table 2, we can see that SiMT-aware strategies based on sentence difficulty and chunk length achieve significant improvements against randomly sampling, where the chunk length-based strategies are the most effective (+0.80 and +0.73 BLEU points for "Align." and "LM", respectively). Besides, the monotonicity-based strategy "Mcity" slightly underperforms the random sampling, especially under high latencies (k =5, 7, 9). The potential reason is "Mcity" prefers short and word-to-word translations, making the sampled synthetic data intuitively easier.

To quantitatively investigate the reason for the slightly worse performance for "Mcity", we visualize the correlations between "Mcity" and other strategies in Figure 2. As shown, the data sampled by the monotonicity-based strategy are significantly different from others. Han et al. (2021) also show that samples chosen by chunk length-based strategy may with poor monotonicity. Given such a huge data gap, it is natural to suspect if there exists a complementary between "Mcity" and the best chunk length-based sampling strategies, e.g., chunk length-based strategy.

Chunk length-based and monotonicity-based strategies complement each other. Based on the above quantitive analysis and suspicion, we combine the chunk length-based strategies and monotonicity-based strategy as follows: 1) sampling monolingual data with the ratio 160% of the original volume according to the chunk length-based strategy, and 2) reranking the sentences with monotonicity-based strategy, and then filter out the extra 60%. As shown in Table 3, we can see that although monotonicity itself does not work well, combining the two gives overall marginal improvements, which is more obvious under low latency, e.g., +0.74 BLEU points improvement on average, indicating the complementary of two types of sampling strategies in difficult scenarios.

Considering the computational complexity of alignment, we will set the LM as the default chunk length-based strategy. Therefore, we leave the combination of LM-based chunk length strategy and monotonicity-based strategy as the default of our method in the following experiments.

Main Results

Figure 3 lists the results on the En⇒Zh and En⇒Ja benchmarks, with *average-lagging* (Ma et al. 2019) being the latency metric. Encouragingly, the conclusions in the empirical findings hold across language pairs, significantly outperforming the random sampling baseline by +0.84 and +0.60 BLEU points, respectively. This demonstrates the effectiveness and universality of our proposed approach. Notably, our data-level approaches neither modify model structure nor add extra training objectives, thus not changing the latency and maintaining the intrinsic advantages of SiMT models. The main side effect of our approach is the increased inference time for building distilled data with sampled monolingual sentences. Fortunately, the cost is once-for-all, and the distilled synthetic data can be flexibly reused. Given the considerable and consistent SiMT improvement, the above cost is acceptable.

Analysis

In this section, we provide quantitative statistics and qualitative cases to show the superiority of our sampling strategy against random sampling.

Similar to full-sentence NMT, SiMT also suffers from hallucination problem (Lee et al. 2018; Chen et al. 2021), generating fluent but inadequate translations, which is caused by overconfidence of the language modeling (Miao et al. 2021). In SiMT, due to the incomplete source sentence, the contribution of source information in prediction

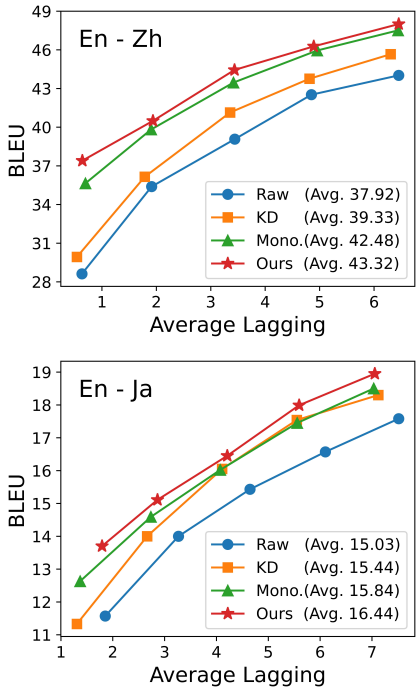

Figure 3: Main results on the En⇒Zh (up) and En⇒Ja (down) benchmarks. Each line represents a system, and the 5 nodes correspond to different wait-k settings ($k = 1, 3, 5, 7, 9$). "Raw" and "KD" represent the systems trained on the original and distilled parallel data, respectively. "Mono." and "Ours" demonstrate using monolingual data with the random sampling strategy and our proposed best strategy, respectively.

is further reduced, resulting in a more serious hallucination problem (Chen et al. 2021). We argue that our strategy is beneficial in avoiding hallucinations, thereby improving the translation quality.

Referring to Chen et al. (2021), we use the hallucination rate of hypotheses to evaluate the generation quality, named *GHall*. In more detail, a target word $\hat{y}_j$ is a *hallucination* if it can not be aligned to any source word it can see currently. Formally, based on word alignment $\mathcal{A}$, whether target word $\hat{y}_j$ is a hallucination is:

$$H(j, \mathcal{A}) = \mathbb{1}[\{(i, j) \in \mathcal{A} \mid i \geq j + k\} = \varnothing]$$

The hallucination rate *GHall* is further defined as:

$$\text{GHall}(\boldsymbol{x}, \hat{\boldsymbol{y}}, \mathcal{A}) = \frac{1}{|\hat{\boldsymbol{y}}|} \sum_{j=1}^{|\hat{\boldsymbol{y}}|} H(j, \mathcal{A})$$

We use the same metric as the monotonicity-based strategy to evaluate the monotonicity of the training set averaged over $k \in 1, 3, 5, 7, 9$, named *TAnti*, and the same metric as the chunk length-based strategy based on alignment to evaluate the average length of the training set (*TCnk*) and generations (*GCnk*).

We first quantitatively analyze how our method affects the constitution of the training data, thereby reducing the translation hallucinations and chunk lengths in Table 4. The anticipation rate and the averaged chunk length of the training data are substantially reduced, leading to a lower hallucination rate and shorter chunks during generation. In addition, we give an example under wait-3 policy in Figure 4 to confirm our claim. The random sampling strategy generates an unwarranted guess at the speaker "NASA says," and mistranslates the phrase "on corals" at the end, while ours perfectly avoids these problems. The above quantitative statistics and qualitative examples demonstrate that our sampling strategy improves the translation against random sampling by reducing the critical issue in SiMT – hallucination.

		TAnti	*TCnk*	*GHall*	*GCnk*
EnZh	Rand.	23.92%	1.11	10.69%	1.11
	Ours	**13.86%**	**1.01**	**8.16%**	**1.08**
EnJa	Rand.	16.47%	1.10	6.91%	1.13
	Ours	**8.30%**	**1.02**	**3.08%**	**1.07**

Table 4: Statistics of monotonicity "*TAnti*" and chunk length "*TCnk*" in monolingual training data, and hallucinations "*GHall*" and chunk length "*GCnk*" in generations.

Scalability Discussion of Our Approach

In this section, we discuss potential directions to further enhance our scalable method to make SiMT a practical translation system by making the most of the 1) monolingual data, 2) larger teacher, and 3) raw bilingual data.

Our strategy performs well with more monolingual data. One strength of using monolingual data is the potential to exploit the scaling ability to further improve translation performance (Edunov et al. 2018; Ding et al. 2022). To validate our claim, we scale the size of monolingual data by $\{\times 3, \times 5, \times 10\}$ and report the performance of random sampling and ours in Table 5. As seen, enlarging the monolingual data consistently improves the BLEU scores, and with scaling factor increases, our strategy achieves higher performance against random ones, e.g., +1.05 BLEU points under 1:10. Besides, the hallucination rate "GHall" and chunk length "GCnk" indicate that ours consistently better than that of random sampling, which validates our claim.

Our strategy performs well with a better teacher. One may expect that augmenting the capacity of the teacher model for our method obtains further improvement. To verify the hypothesis, we employ a larger capacity framework as the teacher, i.e., Transformer-BIG. As shown in Table 6, we see that a larger teacher framework with better translation quality (51.86 vs. 48.55) indeed transfers rich knowledge to the student, further improving the student under all latency settings (+0.56 BLEU points on average).

Our strategy performs well with raw bilingual data. Previous experiments in our study make the combination of distilled bilingual data and synthetic data generated by strategically selected monolingual data as default. Although it has shown significantly better performance against the random sampling strategy, all the training data used to train

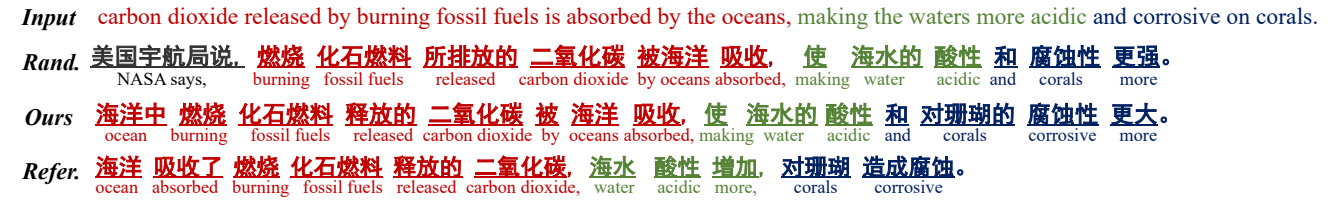

Figure 4: Translation examples of models trained with random "Rand." and our "Ours" monolingual data sampling strategies under the wait-3 policy. "Refer." means the reference. Words without color are hallucinations.

Scale	Strategy	*wait-1*	*wait-3*	*wait-5*	*wait-7*	*wait-9*	*Avg.*	Δ	*GHall*	*GCnk*
1:1	Rand.	35.64	39.82	43.46	45.95	47.51	42.48		10.69%	1.11
	Ours	**37.40**	**40.49**	**44.44**	**46.27**	**48.00**	**43.32**	**+0.84**	**8.16%**	**1.08**
1:3	Rand.	33.79	39.26	43.48	**46.27**	**47.84**	42.13		11.57%	1.13
	Ours	**36.75**	**41.04**	**44.23**	45.99	47.30	**43.06**	**+0.93**	**7.30%**	**1.09**
1:5	Rand.	35.45	39.85	43.26	46.14	**47.70**	42.48		10.79%	1.12
	Ours	**37.35**	**41.40**	**44.65**	**46.35**	47.46	**43.44**	**+0.96**	**6.60%**	**1.07**
1:10	Rand.	34.81	40.54	43.73	45.93	48.02	42.61		10.52%	1.12
	Ours	**37.33**	**42.25**	**44.00**	**46.62**	**48.09**	**43.66**	**+1.05**	**7.26%**	**1.06**

Table 5: Comparison between random sampling "Rand." and "Ours" when scaling up the monolingual data on En⇒Zh. "Scale" refers to the proportion of distilled bilingual data and monolingual data. For translation quality, we report BLEU scores ("wait-k" and "avg." ⇑). For fine-grained evaluation, we report the hallucination rate "*GHall*" (⇓) and chunk length "*GCnk*" (⇑) proposed above. We train all models with the same training steps.

	Teacher	BASE: 48.55	BIG: 51.86	Δ
Student	*wait-1*	37.40	**38.22**	**+0.82**
	wait-3	40.49	**41.84**	**+1.35**
	wait-5	44.44	**44.65**	**+0.21**
	wait-7	46.27	**46.35**	**+0.08**
	wait-9	48.00	**48.34**	**+0.34**
	Avg.	43.32	**43.88**	**+0.56**

Table 6: Augmenting the teacher by employing the teacher with a large model capacity (BIG) on En⇒Zh.

	KD Para. +Mono.	Raw Para. +Mono.	Δ
wait-1	37.40	**38.13**	**+0.73**
wait-3	40.49	**41.72**	**+1.23**
wait-5	**44.44**	44.38	-0.06
wait-7	46.27	**46.61**	**+0.34**
wait-9	**48.00**	47.82	-0.18
Avg.	43.32	**43.73**	**+0.41**

Table 7: Replacing the distilled bilingual data ("KD Para.+") with the raw bilingual data ("Raw Para.+") in our strategy on En⇒Zh, where "KD Para.+ Mono." is the default setting in the previous experiments.

the final SiMT model only utilize the distilled (or synthetic) target-side data, which may lose some long-tailed information in the raw bilingual data (Ding et al. 2021a,b). To verify that the raw bilingual data can further complement our monolingual strategy, we replace the distilled bilingual data with the raw one and report the results in Table 7. We can observe that our strategy performs well with raw bilingual data (+0.41 BLEU points), and the improvements mainly come from the low-latency settings, e.g., +0.73 and +1.23 BLEU points for wait-1 and -3, respectively.

Conclusion

In this work, we first empirically validate the effectiveness of using monolingual data for SiMT. Then, we propose a simple, effective, and scalable monolingual data sampling strategy, considering both the chunk length and monotonicity. Extensive experiments show that our method achieves significant and consistent improvements compared to the random sampling strategy. Analyses verify that our strategy improves the translation quality by alleviating the key problems of SiMT, e.g., the hallucination problem. Furthermore, our method has appealing expandability and can be further enhanced by 1) enlarging the scale of monolingual data, 2) augmenting the capacity of the teacher, and 3) using the raw bilingual data.

Future directions include 1) validating the effectiveness of our data-level method upon advanced SiMT model (Anonymous 2023) and decoding policies (Zhang et al. 2020; Zhang and Feng 2022); and 2) investigating the complementarity (Liu et al. 2021b) between our proposed semi-supervised learning based method and the powerful pre-trained models (Liu et al. 2020b; Zan et al. 2022) in SiMT.

Acknowledgments

We thank the anonymous reviewers for their thorough review and valuable feedback. Liang and Dacheng were supported by the Major Science and Technology Innovation 2030 "Brain Science and Brain-like Research" key project (No. 2021ZD0201405). Xuebo was supported in part by the National Natural Science Foundation of China (Grant No. 62206076 and 62276077) and Shenzhen College Stability Support Plan (Grant No. GXWD20220811173340003 and GXWD20220817123150002).

References

Anonymous. 2023. Hidden Markov Transformer for Simultaneous Machine Translation. In *Submitted to ICLR*.

Arivazhagan, N.; Cherry, C.; Macherey, W.; Chiu, C.-C.; Yavuz, S.; Pang, R.; Li, W.; and Raffel, C. 2019. Monotonic Infinite Lookback Attention for Simultaneous Machine Translation. In *ACL*.

Chen, J.; and Zhang, J. 2019. *Machine Translation: 14th China Workshop, CWMT 2018, Wuyishan, China, October 25-26, 2018, Proceedings*.

Chen, J.; Zheng, R.; Kita, A.; Ma, M.; and Huang, L. 2021. Improving Simultaneous Translation by Incorporating Pseudo-References with Fewer Reorderings. In *EMNLP*.

Chiang, D. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*.

Ding, L.; and Tao, D. 2021. The USYD-JD Speech Translation System for IWSLT2021. In *IWSLT*.

Ding, L.; Wang, L.; Liu, X.; Wong, D. F.; Tao, D.; and Tu, Z. 2021a. Rejuvenating Low-Frequency Words: Making the Most of Parallel Data in Non-Autoregressive Translation. In *ACL*.

Ding, L.; Wang, L.; Liu, X.; Wong, D. F.; Tao, D.; and Tu, Z. 2021b. Understanding and Improving Lexical Choice in Non-Autoregressive Translation. In *ICLR*.

Ding, L.; Wang, L.; Shi, S.; Tao, D.; and Tu, Z. 2022. Redistributing Low-Frequency Words: Making the Most of Monolingual Data in Non-Autoregressive Translation. In *ACL*.

Domhan, T.; and Hieber, F. 2017. Using Target-side Monolingual Data for Neural Machine Translation through Multi-task Learning. In *EMNLP*.

Dyer, C.; Chahuneau, V.; and Smith, N. A. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *NAACL*.

Edunov, S.; Ott, M.; Auli, M.; and Grangier, D. 2018. Understanding Back-Translation at Scale. In *EMNLP*.

Fadaee, M.; and Monz, C. 2018. Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation. In *EMNLP*.

Fukuda, R.; Oka, Y.; Kano, Y.; and et al. 2021. NAIST English-to-Japanese Simultaneous Translation System for IWSLT 2021 Simultaneous Text-to-text Task. In *IWSLT*.

Gu, J.; Neubig, G.; Cho, K.; and Li, V. O. K. 2017. Learning to Translate in Real-time with Neural Machine Translation. In *EACL*.

Han, H.; Ahn, S.; Choi, Y.; Chung, I.; Kim, S.; and Cho, K. 2021. Monotonic Simultaneous Translation with Chunk-wise Reordering and Refinement. In *EMNLP*.

He, H.; Boyd-Graber, J.; and Daumé III, H. 2016. Interpretese vs. Translationese: The Uniqueness of Human Strategies in Simultaneous Interpretation. In *NAACL*.

He, H.; Grissom II, A.; Morgan, J.; Boyd-Graber, J. L.; and Daumé III, H. 2015. Syntax-Based Rewriting for Simultaneous Machine Translation. In *EMNLP*.

Heafield, K.; Pouzyrevsky, I.; Clark, J. H.; and Koehn, P. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *ACL*.

Jiao, W.; Wang, X.; Tu, Z.; Shi, S.; Lyu, M. R.; and King, I. 2021. Self-Training Sampling with Monolingual Data Uncertainty for Neural Machine Translation. In *ACL*.

Kim, Y.; and Rush, A. M. 2016. Sequence-Level Knowledge Distillation. In *EMNLP*.

Koehn, P.; and Knowles, R. 2017. Six Challenges for Neural Machine Translation. In *ACL*.

Kudo, T.; and Richardson, J. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *EMNLP*.

Lee, K.; Firat, O.; Agarwal, A.; Fannjiang, C.; and Sussillo, D. 2018. Hallucinations in Neural Machine Translation. In *IRASL@NeurIPS*.

Liu, D.; Du, M.; Li, X.; Hu, Y.; and Dai, L. 2021a. The USTC-NELSLIP Systems for Simultaneous Speech Translation Task at IWSLT 2021. In *IWSLT*.

Liu, X.; Lai, H.; Wong, D. F.; and Chao, L. S. 2020a. Norm-Based Curriculum Learning for Neural Machine Translation. In *ACL*.

Liu, X.; Wang, L.; Wong, D. F.; Ding, L.; Chao, L. S.; Shi, S.; and Tu, Z. 2021b. On the Complementarity between Pre-Training and Back-Translation for Neural Machine Translation. In *Findings of EMNLP*.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020b. Multilingual Denoising Pre-training for Neural Machine Translation. *TACL*.

Ma, M.; Huang, L.; Xiong, H.; Zheng, R.; Liu, K.; Zheng, B.; Zhang, C.; He, Z.; Liu, H.; Li, X.; Wu, H.; and Wang, H. 2019. STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency Using Prefix-to-Prefix Framework. In *ACL*.

Ma, X.; Dousti, M. J.; Wang, C.; Gu, J.; and Pino, J. M. 2020. SIMULEVAL: An Evaluation Toolkit for Simultaneous Translation. In *EMNLP*.

Miao, M.; Meng, F.; Liu, Y.; Zhou, X.-H.; and Zhou, J. 2021. Prevent the Language Model from Being Overconfident in Neural Machine Translation. In *ACL*.

Moore, R. C.; and Lewis, W. D. 2010. Intelligent Selection of Language Model Training Data. In *ACL*.

Morishita, M.; Suzuki, J.; and Nagata, M. 2020. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In *LREC*.

Platanios, E. A.; Stretcu, O.; Neubig, G.; Póczos, B.; and Mitchell, T. M. 2019. Competence-Based Curriculum Learning for Neural Machine Translation. In *NAACL*.

Post, M. 2018. A Call for Clarity in Reporting BLEU Scores. In *WMT*.

Ren, Y.; Liu, J.; Tan, X.; Zhang, C.; Qin, T.; Zhao, Z.; and Liu, T.-Y. 2020. SimulSpeech: End-to-End Simultaneous Speech to Text Translation. In *ACL*.

Schwenk, H.; Chaudhary, V.; Sun, S.; Gong, H.; and Guzmán, F. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *EACL*.

Sen, S.; Germann, U.; and Haddow, B. 2021. The University of Edinburgh's Submission to the IWSLT21 Simultaneous Translation Task. In *IWSLT*.

Sennrich, R.; Haddow, B.; and Birch, A. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *ACL*.

Sennrich, R.; Haddow, B.; and Birch, A. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; and et al. 2017. Attention Is All You Need. In *NeurIPS*.

Wu, L.; Wang, Y.; Xia, Y.; Qin, T.; Lai, J.; and Liu, T.-Y. 2019. Exploiting Monolingual Data at Scale for Neural Machine Translation. In *EMNLP*.

Zan, C.; Peng, K.; Ding, L.; and et al. 2022. Vega-MT: The JD Explore Academy Translation System for WMT22. In *WMT*.

Zhang, J.; and Zong, C. 2016. Exploiting Source-side Monolingual Data in Neural Machine Translation. In *EMNLP*.

Zhang, R.; Zhang, C.; He, Z.; Wu, H.; and Wang, H. 2020. Learning Adaptive Segmentation Policy for Simultaneous Translation. In *EMNLP*.

Zhang, S.; and Feng, Y. 2022. Information-Transport-based Policy for Simultaneous Translation. In *EMNLP*.

Zhang, S.; Feng, Y.; and Li, L. 2021. Future-Guided Incremental Transformer for Simultaneous Translation. In *AAAI*.

Zhao, C.; Liu, Z.; Tong, J.; Wang, T.; Wang, M.; Ye, R.; Dong, Q.; Cao, J.; and Li, L. 2021. The Volctrans Neural Speech Translation System for IWSLT 2021. In *IWSLT*.

Zheng, B.; Liu, K.; Zheng, R.; Ma, M.; Liu, H.; and Huang, L. 2020. Simultaneous Translation Policies: From Fixed to Adaptive. In *ACL*.

Zhou, C.; Gu, J.; and Neubig, G. 2020. Understanding Knowledge Distillation in Non-autoregressive Machine Translation. In *ICLR*.

Zhou, J.; and Keung, P. 2020. Improving Non-autoregressive Neural Machine Translation with Monolingual Data. In *ACL*.

Zhu, X.; and Goldberg, A. B. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1): 1–130.