Improving Simultaneous Machine Translation with Monolingual Data

Hexuan Deng\textsuperscript{1*}, Liang Ding\textsuperscript{2}, Xuebo Liu\textsuperscript{1†}, Meishan Zhang\textsuperscript{1}, Dacheng Tao\textsuperscript{2}, Min Zhang\textsuperscript{1}

\textsuperscript{1} Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
\textsuperscript{2} JD Explore Academy, JD.com Inc.

\texttt{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$\Rightarrow$Zh and En$\Rightarrow$Ja news domain corpora demonstrate that monolingual data can significantly improve translation quality (e.g., +3.15 BLEU on En$\Rightarrow$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$\Rightarrow$Zh and En$\Rightarrow$Ja. Data and codes can be found at https://github.com/hexuanpeng/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 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.

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$\Rightarrow$Zh and En$\Rightarrow$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.

\textsuperscript{1}Work was done when Hexuan was interning at JD Explore Academy.
\textsuperscript{†}Corresponding author: Xuebo Liu.
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
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 \( x = (x_1, \ldots, x_m) \) as input, and outputs hidden state \( h = (h_1, \ldots, h_m) \). 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:

\[
\hat{y}_t = \arg\max_{y_t} p(y_t \mid x, y_{<t}; \theta) \quad (1)
\]

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:

\[
\hat{y}_t = \arg\max_{y_t} p(y_t \mid x, y_{<t}; \theta) \quad (2)
\]

where \( g_{\text{wait},k}(t) = \min\{k + t - 1, |x|\} \) 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 hand-writing 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 alleviating 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 \( \alpha \) by replacing the sentence length term (or similar item) \(|x|\) with \(|x|^\alpha \) (or \(|x|^{1/\alpha}\)), 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
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 \( \alpha \) 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 \( \ell \), the chunk length-based metric for the sentence is:

\[
S_{\text{chunk}} = \frac{\ell^\alpha}{c}
\]  

(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 \( x_{ik} \sim y_{jk} \), the monotonicity-based metric for the sentence is:

\[
S_{\text{mono}} = \frac{1}{|\mathcal{A}|^{1/\alpha}} \sum_{t=1}^{|\mathcal{A}|} \mathbf{1}[i_t \leq j_t + k]
\]  

(4)

where \( \alpha \) 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 \( \alpha \) 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 \( \mathbf{x} = (x_1, ..., x_m) \), 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:

\[
S_{\text{rarity}} = -\frac{1}{|\mathbf{x}|^{\alpha}} \sum_{i=1}^{|\mathbf{x}|} \log \hat{p}(x_i)
\]  

(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 \( \mathbf{x} = (x_1, ..., x_m) \), 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:

\[
E(y | \mathcal{A}, x_i) = -\sum_{y_j \in \mathcal{A}(x_i)} p(y_j | x_i) \log p(y_j | x_i)
\]

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

\[
S_{\text{uncer}} = \frac{1}{|\mathbf{x}|^{\alpha}} \sum_{i=1}^{|\mathbf{x}|} E(y | \mathcal{A}, x = x_i)
\]  

(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\(^2\) (Mori-shita, Suzuki, and Nagata 2020) and WikiMatrix\(^3\) (Schwenk et al. 2021) as training data, newsdev2020 as the validation data.

\(^1\)http://nlp.nju.edu.cn/cwmt-wmt/
\(^2\)https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/
\(^3\)https://opus.nlpl.eu/WikiMatrix.php
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) 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-sequence Big Trans-

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

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
<th>KD</th>
<th>KD+Mono.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Teacher: 48.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wait-1</td>
<td>28.62</td>
<td>29.93</td>
<td>35.64</td>
</tr>
<tr>
<td>wait-3</td>
<td>35.39</td>
<td>36.15</td>
<td>39.82</td>
</tr>
<tr>
<td>wait-5</td>
<td>39.07</td>
<td>41.14</td>
<td>43.46</td>
</tr>
<tr>
<td>wait-7</td>
<td>42.52</td>
<td>43.76</td>
<td>45.95</td>
</tr>
<tr>
<td>wait-9</td>
<td>44.02</td>
<td>45.66</td>
<td>47.51</td>
</tr>
<tr>
<td>Avg.</td>
<td>37.92</td>
<td>39.33</td>
<td>42.48 (+1.41/-) (+4.56/1.15)</td>
</tr>
</tbody>
</table>

Monolingual Data

We closely follow previous works to randomly select monolingual data from publicly available News Crawl corpus\(^3\) (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\textasciitilde2017 and News Discussions 2014\textasciitilde2017.

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-sequence Big Trans-
The main side effect of our approach is the increased infer-
ence time for building distilled data with sampled monolin-
ual sentences. Fortunately, the cost is once-for-all, and the
distilled synthetic data can be flexibly reused. Given the con-
siderable and consistent SiMT improvement, the above cost
is acceptable.

**Main Results**

Figure 3 lists the results on the En⇒Zh and En⇒Ja bench-
marks, with average-lagging (Ma et al. 2019) being the la-
tency metric. Encouragingly, the conclusions in the empiri-
cal findings hold across language pairs, significantly outper-
forming the random sampling baseline by +0.84 and +0.60
BLEU points, respectively. This demonstrates the effective-
ness 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 infer-
ence time for building distilled data with sampled monolin-
gual sentences. Fortunately, the cost is once-for-all, and the
distilled synthetic data can be flexibly reused. Given the con-
siderable and consistent SiMT improvement, the above cost
is acceptable.

**Analysis**

In this section, we provide quantitative statistics and quali-
tative 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 sen-
tence, the contribution of source information in prediction

---

**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”.

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

**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 ∆.

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

**SiMT-aware sampling strategies do help.** We test the ef-
fects 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 signif-
ificant 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”, respec-
tively). Besides, the monotonicity-based strategy “Mcity”
slightly underperforms the random sampling, especially un-
der high latencies (k = 5, 7, 9). The potential reason is “Mc-
ity” 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 correla-
tions between “Mcity” and other strategies in Figure 2. As
shown, the data sampled by the monotonicity-based strat-
 egy are significantly different from others. Han et al. (2021)
also show that samples chosen by chunk length-based strat-
ey 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 quantitative
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 orig-
inal volume according to the chunk length-based strategy,
and 2) reranking the sentences with monotonicity-based strat-
 egy, and then filter out the extra 60%. As shown in Ta-
ble 3, we can see that although monotonicity itself does not
work well, combining the two gives overall marginal im-
provements, 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 diffi-
cult scenarios.

Considering the computational complexity of alignment,
we will set the LM as the default chunk length-based strat-
 egy. 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.
The hallucination rate can not be aligned to any source word it can see currently. In more detail, a target word translation quality. We use the same metric as the monotonicity-based strategy, respectively.

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 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.

Scaleability 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.

<table>
<thead>
<tr>
<th></th>
<th>Rand.</th>
<th>Ours</th>
<th>Rand.</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAnti</td>
<td>23.92%</td>
<td>13.86%</td>
<td>16.47%</td>
<td>8.30%</td>
</tr>
<tr>
<td>GCnk</td>
<td>1.11</td>
<td>1.01</td>
<td>1.02</td>
<td>1.08</td>
</tr>
<tr>
<td>GHall</td>
<td>10.69%</td>
<td>8.16%</td>
<td>6.91%</td>
<td>3.08%</td>
</tr>
<tr>
<td>GCnk</td>
<td>1.11</td>
<td>1.08</td>
<td>1.13</td>
<td>1.07</td>
</tr>
</tbody>
</table>

Table 4: Statistics of monotonicity “TAnti” and chunk length “TCnk” in monolingual training data, and hallucinations “GHall” and chunk length “GCnk” in generations.
carbon dioxide released by burning fossil fuels is absorbed by the oceans, making the waters more acidic and corrosive on corals.

_**Rand.**_ NASA says, _**Ours**_ ocean burning fossil fuels released carbon dioxide by oceans absorbed, making water acidic and corrosive more.

**Refer.**_ ocean absorption fossil fuels released carbon dioxide by oceans absorbed, making water acidic and corrosive more.

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.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Strategy</th>
<th>wait-1</th>
<th>wait-3</th>
<th>wait-5</th>
<th>wait-7</th>
<th>wait-9</th>
<th>Avg.</th>
<th>∆</th>
<th>GHall</th>
<th>GCnk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1</td>
<td>Rand.</td>
<td>35.64</td>
<td>39.82</td>
<td>43.46</td>
<td>45.95</td>
<td>47.51</td>
<td>42.48</td>
<td>1.11</td>
<td>10.69%</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>37.40</td>
<td>40.49</td>
<td>44.44</td>
<td>46.27</td>
<td>48.00</td>
<td>43.32 +0.84</td>
<td>9.16%</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>1:3</td>
<td>Rand.</td>
<td>33.79</td>
<td>39.26</td>
<td>43.48</td>
<td>46.27</td>
<td>47.84</td>
<td>42.13</td>
<td>1.13</td>
<td>11.57%</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>36.75</td>
<td>41.04</td>
<td>44.23</td>
<td>45.99</td>
<td>47.30</td>
<td>43.06 +0.93</td>
<td>7.30%</td>
<td>1.09</td>
<td></td>
</tr>
<tr>
<td>1:5</td>
<td>Rand.</td>
<td>35.45</td>
<td>39.85</td>
<td>43.26</td>
<td>46.14</td>
<td>47.70</td>
<td>42.48</td>
<td>1.12</td>
<td>10.79%</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>37.35</td>
<td>41.40</td>
<td>44.65</td>
<td>46.35</td>
<td>47.46</td>
<td>43.44 +0.96</td>
<td>6.66%</td>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td>1:10</td>
<td>Rand.</td>
<td>34.81</td>
<td>40.54</td>
<td>43.73</td>
<td>45.93</td>
<td>48.02</td>
<td>42.61</td>
<td>1.12</td>
<td>10.52%</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>37.33</td>
<td>42.25</td>
<td>44.00</td>
<td>46.62</td>
<td>48.09</td>
<td>43.66 +1.05</td>
<td>7.26%</td>
<td>1.06</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Teacher</th>
<th>BASE: 48.55</th>
<th>BIG: 51.86</th>
<th>∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>wait-1</td>
<td>37.40</td>
<td>38.22</td>
<td>+0.82</td>
</tr>
<tr>
<td>wait-3</td>
<td>40.49</td>
<td>41.84</td>
<td>+1.35</td>
</tr>
<tr>
<td>wait-5</td>
<td>44.44</td>
<td>44.65</td>
<td>+0.21</td>
</tr>
<tr>
<td>wait-7</td>
<td>46.27</td>
<td>46.35</td>
<td>+0.08</td>
</tr>
<tr>
<td>wait-9</td>
<td>48.00</td>
<td>48.34</td>
<td>+0.34</td>
</tr>
<tr>
<td>Avg.</td>
<td>43.32</td>
<td>43.88</td>
<td>+0.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>KD Para.</th>
<th>Raw Para.</th>
<th>∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>wait-1</td>
<td>37.40</td>
<td>38.13</td>
</tr>
<tr>
<td>wait-3</td>
<td>40.49</td>
<td>41.72</td>
</tr>
<tr>
<td>wait-5</td>
<td>44.44</td>
<td>44.38</td>
</tr>
<tr>
<td>wait-7</td>
<td>46.27</td>
<td>46.61</td>
</tr>
<tr>
<td>wait-9</td>
<td>48.00</td>
<td>47.82</td>
</tr>
<tr>
<td>Avg.</td>
<td>43.32</td>
<td>43.73</td>
</tr>
</tbody>
</table>

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

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.

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


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


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


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


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


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. 2016b. Neural Machine Translation of Rare Words with Subword Units. In ACL.


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


