Dynamic Data Selection and Weighting for Iterative Back-Translation

Zi-Yi Dou[з] **Antonios Anastasopoulos**[ð,†] **Graham Neubig**[з]
[з] Language Technologies Institute, Carnegie Mellon University
[ð] Department of Computer Science, George Mason University
{zdou,gneubig}@cs.cmu.edu antonis@gmu.edu

Abstract

Back-translation has proven to be an effective method to utilize monolingual data in neural machine translation (NMT), and iteratively conducting back-translation can further improve the model performance. Selecting which monolingual data to back-translate is crucial, as we require that the resulting synthetic data are of high quality *and* reflect the target domain. To achieve these two goals, data selection and weighting strategies have been proposed, with a common practice being to select samples close to the target domain but also dissimilar to the average general-domain text. In this paper, we provide insights into this commonly used approach and generalize it to a dynamic curriculum learning strategy, which is applied to iterative back-translation models. In addition, we propose weighting strategies based on both the current quality of the sentence and its improvement over the previous iteration. We evaluate our models on domain adaptation, low-resource, and high-resource MT settings and on two language pairs. Experimental results demonstrate that our methods achieve improvements of up to 1.8 BLEU points over competitive baselines.[1]

1 Introduction

Back-translation (Sennrich et al., 2016b) is an effective strategy for improving the performance of neural machine translation (NMT) using monolingual data, delivering impressive gains over already competitive NMT models (Edunov et al., 2018). The strategy is simple: given monolingual data in the target language, one can use a translation model in the opposite of the desired translation direction to *back-translate* the monolingual data, effectively synthesizing a parallel dataset, which is in turn used to train the final translation model. Further improvements can be obtained by iteratively repeating this process (Hoang et al., 2018) in both directions.

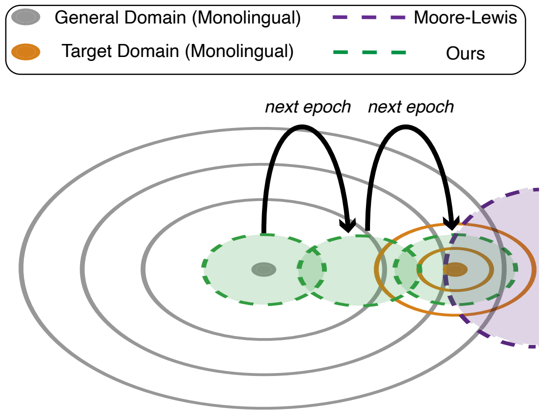

Figure 1: The Moore and Lewis (2010) data selection strategy for domain adaptation constantly selects the same set of sentences which cannot well represent the target domain. Our approach, instead, selects different subsets of sentences at each epoch and we gradually shift from selecting samples from the general-domain distribution to samples from the target distribution.

However, not all monolingual data are equally important. An envisioned downstream application is very often characterized by a unique data distribution. In such cases of domain shift, back-translating target domain data can be an effective strategy (Hu et al., 2019) for obtaining a better in-domain translation model. One common strategy is to select samples that are both (1) close to the target distribution and (2) dissimilar to the average general-domain text (Moore and Lewis, 2010). However, as depicted in Figure 1, this method is not ideal because the second objective could bias towards the selection of sentences far from the center of the target distribution, potentially leading to selecting a non-representative set of sentences.

Even if we could select all in-domain monolin-

†: Work completed while at Carnegie Mellon University.
[1]Code: https://github.com/zdou0830/dynamic_select_weight.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 5894–5904,
November 16–20, 2020.

gual data, the back-translation model has not been trained on in-domain parallel data and thus the back-translated data will be of poor quality. As we demonstrate in the experiments, the quality of the back-translated data can have a large influence on the final model performance.

To achieve the two goals of both selecting target-domain data and back-translating them with high quality, in this paper, we propose a method to combine dynamic data selection with weighting strategies for iterative back-translation. Specifically, the dynamic data selection selects subsets of sentences from a monolingual corpus *at each training epoch*, gradually transitioning from selecting general-domain data to choosing target-domain sentences. The gradual transition ensures that the back-translation model of each iteration can adequately translate the selected sentences, as they are close to the distribution of its current training data. We also assign weights to the back-translated data that reflect their quality, which further reduces the effect of potential noise due to low quality translations. The proposed data selection and weighting strategies are complementary to each other, as the former focuses on domain information while the latter emphasizes the quality of sentences.

We investigate the performance of our methods in domain adaptation, low-resource and high-resource MT settings and on German-English and Lithuanian-English datasets. Our strategies demonstrate improvements of up to 1.8 BLEU points over a competitive iterative back-translation baseline and up to 1.2 BLEU points over the best static data selection strategies. In addition, our analysis reveals that the selected samples can represent the target distribution well and that the weighting strategies are effective in noisy settings.

2 Background: Back-Translation

Back-translation (Sennrich et al., 2016a) has proven to be an effective way of utilizing monolingual data for machine translation. Given a parallel training corpus $\mathbf{D}_{FE}$, we first train a target-to-source machine translation model $\mathbf{M}_{EF}$. Then, we use the pre-trained model $\mathbf{M}_{EF}$ to translate a target language monolingual corpus $\mathbf{D}_E$ to the source language and obtain a synthetic parallel corpus $(\mathbf{D}'_F, \mathbf{D}_E)$. Last, we concatenate back-translated data $(\mathbf{D}'_F, \mathbf{D}_E)$ with the original parallel corpus $\mathbf{D}_{FE}$ to train a source-to-target model $\mathbf{M}_{FE}$.

The success of back-translation has motivated researchers to investigate and extend the method (He et al., 2016; Zheng et al., 2020). Hoang et al. (2018) propose to use iterative back-translation and achieve improvements over previous state-of-the-art models. As shown in Algorithm 1, at each training step, a batch of monolingual sentences is sampled from one language and back-translated to the other language. The back-translated data is utilized to train the model in the other direction. The process is repeated in both directions.

Algorithm 1 Iterative Back-Translation

Input: Monolingual corpora $\mathbf{D}_F$ and $\mathbf{D}_E$
Output: Translation models $\mathbf{M}_{FE}$ and $\mathbf{M}_{EF}$
  while $\mathbf{M}_{FE}$ and $\mathbf{M}_{EF}$ have not converged **do**
    for all batches $(\mathbf{B}_F, \mathbf{B}_E)$ in $(\mathbf{D}_F, \mathbf{D}_E)$ **do**
      Translate $\mathbf{B}_F$ into $\mathbf{B}'_E$ using $\mathbf{M}_{FE}$
      Translate $\mathbf{B}_E$ into $\mathbf{B}'_F$ using $\mathbf{M}_{EF}$
      Train $\mathbf{M}_{FE}$ with $(\mathbf{B}'_F, \mathbf{B}_E)$
      Train $\mathbf{M}_{EF}$ with $(\mathbf{B}'_E, \mathbf{B}_F)$
    end for
  end while

3 Methods

In our setting, we are given two MT models $\mathbf{M}_{FE}$ and $\mathbf{M}_{EF}$ pretrained on parallel data $\mathbf{D}_{FE}$, and both source and target monolingual corpora $\mathbf{D}_F$ and $\mathbf{D}_E$. The goal is to select and weight samples from the two monolingual corpora for back-translation, in order to best improve the performance of the two translation models.

3.1 Data Selection Strategies

We first describe a commonly used *static* selection strategy, and then illustrate our *dynamic* approach.

3.1.1 The Moore and Lewis (2010) Method

A common approach for data selection is the Moore and Lewis (2010) method (and extensions, e.g. Axelrod et al. (2011); Duh et al. (2013); Santamaría and Axelrod (2019)), which computes the language model cross-entropy difference for each sentence s in a monolingual corpus:

$$score(\mathbf{s}) = H_{LM_{in}}(\mathbf{s}) - H_{LM_{gen}}(\mathbf{s}), \quad (1)$$

where $H_{LM_{in}}(\mathbf{s})$ and $H_{LM_{gen}}(\mathbf{s})$ represent the cross-entropy scores of s measured with an in-domain and a general-domain language model (LM) respectively. Sentences with the highest scores will be selected for training. Typically, the in-domain language model LM_{in} is trained with

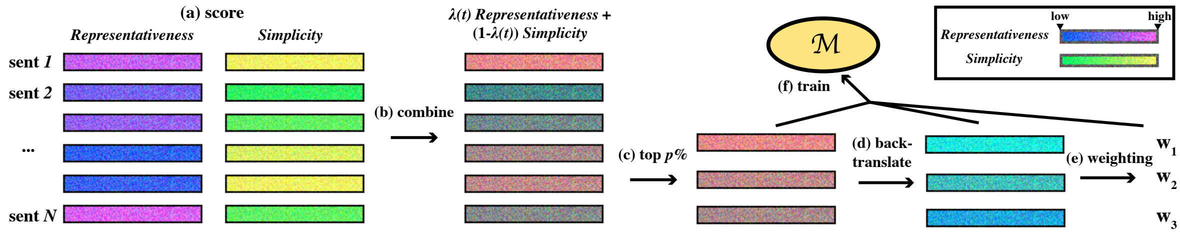

Figure 2: Main procedure of our algorithm. We first compute the representative and simplicity scores for all the monolingual sentences (a). At each training epoch t, we combine the two scores (b) and select the top $p\%$ monolingual sentences (c). After back-translating the selected sentences from the source side to the target side (d), we then perform data weighting on the back-translated samples (e) and train the model with the weighted back-translated sentences (f).

a small set of sentences in the target domain and LM_{gen} is trained with all data available.

3.1.2 Our Two Scoring Criteria

Instead of static data selection, we propose a new curriculum strategy for iterative back-translation. Specifically, we measure both **representativeness**, *i.e.* how well the sentence represents the target distribution, and **simplicity**, *i.e.* how well the MT models can translate the sentence, of each sentence s in the monolingual corpus. First, we select the most simple samples for back-translation to ensure the quality of the back-translated data. As the training progresses, the model will become better at translating in-domain sentences, and we will then shift to choosing more representative examples.

Formally, at each epoch t, we rank the corpus according to

$$\text{score}(\mathbf{s}) = \lambda(t)\text{repr}(\mathbf{s}) + (1 - \lambda(t))\text{simp}(\mathbf{s}), \quad (2)$$

where $\text{repr}(\mathbf{s})$ and $\text{simp}(\mathbf{s})$ denote the representativeness and simplicity of sentence s respectively, which will be dicussed in the following sections. The term $\lambda(t)$ balances between the two criteria and is a function of the current epoch t.

We adopt the square-root growing function for λ (Platanios et al., 2019) and set

$$\lambda(t) = \min(1, \sqrt{t\frac{1 - c_0^2}{T} + c_0^2}), \qquad (3)$$

where c_0 is the initial value and T denotes the time after which we solely select representative samples. λ increases relatively quickly at first and then its acceleration will be gradually decreased as the training progresses, which is suitable for our task as at first the sentences are relatively simple and thus we will not need much time on those sentences.

Connections to Moore and Lewis (2010). Our proposed criteria generalize Moore and Lewis (2010). The first term of Equation 1, namely $H_{LM_{in}}(\mathbf{s})$, measures the **representativeness** of data because the in-domain LM assigns low entropy to sentences that appear frequently in the target domain. The second term $H_{LM_{gen}}(\mathbf{s})$, on the other hand, measures the **simplicity** of the sentences. If $H_{LM_{gen}}(\mathbf{s})$ is high, it is likely that some n-grams of the sentence s appear frequently in the parallel training data $\mathbf{D}_{FE}$, indicating that the MT models will likely translate the sentence well. In other words, the sentence s can provide limited additional information if $H_{LM_{gen}}(\mathbf{s})$ is high. Therefore, one can view Moore and Lewis (2010) as selecting the most representative and difficult sentences.

3.1.3 Representativeness Metrics

We propose three approaches to measure the sentence representativeness.

In-Domain Language Model Cross-Entropy (LM-in). As in Axelrod et al. (2011); Duh et al. (2013), we can use $H_{LM_{in}}$ to measure the representativeness of the instances. Concretely, we train a language model LM_{in} with in-domain monolingual data and compute the score $\frac{1}{|\mathbf{s}|}\sum_{t=1}^{|\mathbf{s}|}\log P_{LM_{in}}(s_t|s_{<t})$ for each sentence s.

TF-IDF Scores (TF-IDF). TF-IDF score is another criterion for data selection (Kirchhoff and Bilmes, 2014). For each sentence s, one can compute its term frequency and inverse document frequency for each word. We can thus obtain the TF-IDF vector and calculate the cosine similarity between the TF-IDF vectors of s and each sentence $\mathbf{s}_{in}$ in a small in-domain dataset, and treat the maximum value as its representativeness score.

BERT Representation Similarities (BERT). BERT (Devlin et al., 2019) has proven to be effective for sentence representation learning. Following the conclusion of Pires et al. (2019), we feed each sentence to the multilingual BERT model and average the hidden states for all the input tokens except [CLS] and [SEP] at the eighth layer to obtain the sentence representation. We then compute the cosine similarity between representations of sentence $\mathbf{s}$ in the monolingual corpus and each sentence $\mathbf{s}_{in}$ in a small in-domain set, and the maximum value is treated as the representativeness score.

3.1.4 Simplicity Metrics

In our experiments, we study two metrics for measuring the simplicity of sentences. Note that in the field of quality estimation for MT (Specia et al., 2010; Fonseca et al., 2019), researchers have proposed several existing techniques to estimate the simplicity of sentences (Turchi et al., 2014; Specia et al., 2015; Shah et al., 2015; Kim and Lee, 2016; Kepler et al., 2019; Zhou et al., 2019; Hou et al., 2019), and here we select a few representative approaches.

General-Domain Language Model Cross-Entropy (LM-gen). We train a language model LM_{gen} with the one side of the parallel training data $\mathbf{D}_{FE}$. Then, for each sentence $\mathbf{s}$ we compute the score $\frac{1}{|\mathbf{s}|}\sum_{t=1}^{|\mathbf{s}|}\log P_{LM_{gen}}(s_t|s_{<t})$.

Round-Trip BLEU (R-BLEU). Given two pretrained MT models $\mathbf{M}_{FE}$ and $\mathbf{M}_{EF}$, round-trip translation first translates a sentence $\mathbf{s}$ into another language using $\mathbf{M}_{FE}$ and then back-translates the result using $\mathbf{M}_{EF}$, obtaining the reconstructed sentence $\mathbf{s}'$. The BLEU score between $\mathbf{s}$ and $\mathbf{s}'$ is treated as our simplicity metric. Similar ideas have been applied to filter sentences of low quality (Imankulova et al., 2017).

For both the representativeness and simplicity scores, it should be noted that they are separately normalized to [0, 1], using the equation $\frac{score(\mathbf{s})-score_{min}}{score_{max}-score_{min}}$, where $score_{max}$ and $score_{min}$ are the maximum and minimum scores.

3.2 Weighting Strategies

Next, we illustrate how we perform data weighting on the back-translated data.

3.2.1 Measuring the Current Quality

As general translation models could perform poorly on the in-domain data, we need ways to measure the current quality of the back-translated sentences in order to down-weight examples of poor quality.

Encoder Representation Similarities (Enc). We feed the source sentence $\mathbf{x}$ and the target sentence $\mathbf{y}$ to the encoders of $\mathbf{M}_{FE}$ and $\mathbf{M}_{EF}$ respectively, and average the hidden states at the final layer to obtain the representations $enc_{FE}(\mathbf{x})$ and $enc_{EF}(\mathbf{y})$. The cosine similarity between them is treated as the quality metric.

Agreement Between Forward and Backward Models (Agree). Inspired by Junczys-Dowmunt (2018), the second approach utilizes the agreement of the two translation models. For each sentence pair $(\mathbf{x}, \mathbf{y})$, we compute the conditional probability $H_{FE}(\mathbf{y}|\mathbf{x})$ and $H_{EF}(\mathbf{x}|\mathbf{y})$, then exponentiate the absolute value between them $\exp(-(|H_{FE}(\mathbf{y}|\mathbf{x}) - H_{EF}(\mathbf{x}|\mathbf{y})|))$. Intuitively, the back-translated sentences are of poor quality if there are huge disagreements between the two models.

3.2.2 Measuring Quality Improvements

In domain adaptation, it is natural that at first the in-domain sentences are poorly translated. As training progresses, however, the quality should be improved. We therefore propose a metric to measure the improvement in translation quality and combine it with the current quality metric, in order to encourage the inclusion of in-domain sentences where the translation qualities have improved.

Specifically, every time we obtain the quality score of sentence $\mathbf{s}$, we store it, then the next time we come across the same sentence, we can compare the new quality score with the previous one:

$$\text{Imp}(\mathbf{s}) = clip(\frac{\text{current\_quality}(\mathbf{s})}{\text{previous\_quality}(\mathbf{s})}, w_{low}, w_{high}),$$

where the clipping function limits the weights to a reasonable range. We set (w_{low}, w_{high}) to $(\frac{1}{2}, 2)$.

3.3 Overall Algorithm: Combining Curriculum and Weighting Strategies

Our final algorithm is shown in Figure 2. At each epoch, we compute the score for each sentence in monolingual corpora using Equation 2 and select the top $p\%$ of sentences, where p is a hyper-parameter. Afterwards, we perform back-translation and data weighting on the selected data, then use the back-translated data to train the translation model. The process will be repeated iteratively for both directions, with λ increased at each training epoch.

Method	WMT			
	LAW		MED	
	de-en	en-de	de-en	en-de
Baseline				
Base	31.25	24.44	34.43	26.59
Back	35.90	26.33	42.42	33.98
Ite-Back	37.69	27.81	44.08	35.65
Zhang et al. (2019)	37.70	27.87	44.25	36.01
Best Selection				
TF-IDF	38.26*	28.35*	44.26	35.82
Best Curriculum				
TF-IDF + R-BLEU	**39.11***	28.93*	44.91*	36.19*
Best Weighting				
Enc	38.20*	28.15*	44.28*	35.52
Enc-Imp	38.13*	27.97	44.46*	35.77
Best Curriculum + Best Weighting				
Curri+Enc	38.87	**29.04**	**45.46***	36.34
Curri+Enc-Imp	38.75	28.89	**45.46***	**36.45***

Table 1: Translation accuracy (BLEU (Papineni et al., 2002)) in the domain adaptation setting. The first and second rows list source and target domains respectively. The third row lists the translation directions. We report the best-performing models of only using selection strategies (*"Best Selection")*, only using curriculum strategies (*"Best Curriculum"*), only using weighting strategies (*"Best Weighting"*) and using both the best curriculum and weighting strategies (*"Best Weighting + Best Weighting"*). "Enc-Imp" indicates both the encoder representation similarities and the quality improvement metrics are used for weighting. The highest scores are in **bold** and $*$ indicates statistical significance compared with the best baseline ($p < 0.05$).

Method	WMT			
	LAW		MED	
	de-en	en-de	de-en	en-de
Baseline				
Ite-Back	37.69	27.81	44.08	35.65
Selection				
BERT	37.84	28.12	44.17	35.68
LM-diff	37.91	27.77	**44.59**	**36.00**
LM-in	38.23	28.29	44.25	34.98
TF-IDF	**38.26**	**28.35**	44.26	35.82
Weighting				
Enc	**38.20**	**28.15**	44.28	35.52
Enc-Imp	38.13	27.97	**44.46**	**35.77**
Agree	37.41	27.70	44.04	35.70
Agree-Imp	37.42	27.78	44.30	35.37
Curriculum				
LM-in+ LM-gen	38.26	28.51	44.68	34.90
TF-IDF + LM-gen	38.67	28.67	44.90	35.49
TF-IDF + R-BLEU	***39.11***	***28.93***	***44.91***	***36.19***

Table 2: Comparisons of different metrics in domain adaptation. The highest scores in each section are in **bold** and the overall highest scores are in ***bold italic***.

4 Experiments on Domain Adaptation

We first conduct experiments in the domain adaptation setting, where we adapt models from a general domain to a specific domain.

4.1 Setup

Datasets. We first train the translation models with (general-domain) WMT-14 German-English dataset, consisting of about 4.5M training sentences, then perform iterative back-translation with (in-domain) law or medical OPUS monolingual data (Tiedemann, 2012). We de-duplicate the law and medical parallel training data, divide them into two halves and obtain 250K and 200K comparable yet non-parallel sentences respectively in both languages to obtain the monolingual corpora. The development and test sets contain 2K sentences in each domain. Byte-pair encoding (Sennrich et al., 2016b) is applied with 32K merge operations. The general-domain and in-domain language models are trained on the WMT training data and the OPUS monolingual data respectively. The OPUS development sets are used to compute the TF-IDF and BERT representativeness scores.

Models. We implement our approaches upon the Transformer (Vaswani et al., 2017). Both the encoder and decoder consist of 6 layers and the hidden size is set to 512. For the translation models, weights of the top 4 layers of the encoders and bottom 4 layers of the decoders are shared between forward and backward models. We also tie the source and target word embeddings. We build 5-gram language models with modified Kneser-Ney smoothing using KenLM (Heafield, 2011).

Hyper-Parameters. c_0 and T in Equation 3 are set to 0.1 and 5. We select 30% of the sentences with the highest score at each epoch for our curriculum methods and 50% of the sentences for the static data selection baselines.

4.2 Results

We compare our dynamic curriculum and weighting methods with three baselines: the iterative back-translation baseline, a baseline trained with only data selection strategies, a baseline trained with only data weighting strategies. The results with the best-performing representativeness and simplicity metrics (TF-IDF and R-BLEU, respectively) in the

Method	de-en	en-de	Avg. Δ
Ite-Sampling	34.93	26.72	-
Ite-Sampling + Enc	35.67	27.76	+0.89
Ite-Greedy	37.69	27.81	-
Ite-Greedy+Enc	38.20	28.15	+0.43
Ite-Beam	37.53	28.25	-
Ite-Beam+Enc	37.76	28.25	+0.12

Table 3: Noise in back-translated data can degrade the model performance and our weighting strategies (Enc) benefit the most in noisy settings.

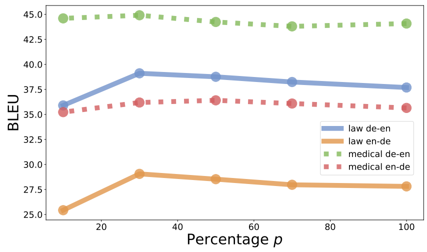

Figure 3: While the model is relatively robust to the number of selected sentences at each epoch, selecting too many or too few sentences can be harmful.

domain adaptation setting are listed in Table 1.

Iterative Back-Translation. The iterative back-translation method is rather competitive, as it improves over the unadapted baseline by 9.6 BLEU and simple back-translation by 1.8 BLEU points.

Selection Strategies. We can see from the table that the best-performing selection strategies, namely selecting sentences with high TF-IDF scores, is generally effective and can improve the baseline by about 0.5 BLEU points.

Curriculum and Weighting Strategies. Both our curriculum and weighting strategies outperform the unadapted and the iterative back-translation models, as well as the curriculum method proposed in Zhang et al. (2019), with our curriculum learning method achieving better performance and improving the strong iterative back-translation baseline by 1.1 BLEU points. Combining curriculum and weighting methods can further improve the performance by up to 0.5 BLEU points, demonstrating the two strategies are complementary to each other.

4.3 Choices of Metrics

We examine different choices of representativeness and simplicity metrics. The performance of different models is listed in Table 2.

Representativeness Metrics. All data selection strategies outperform the baseline, with TF-IDF, LM-diff, and BERT metrics exhibiting fairly robust performance in all settings. Due to its simplicity, we choose TF-IDF for experiments where a good in-domain development set is available.

Data Weighting Strategies. The agreement-based weighting method ("Agree") performs slightly worse than the encoder-similarity weighting strategy ("Enc"), probably because the two languages are similar and thus encoders with shared parameters can accurately measure the data quality.

Curriculum Strategies. Table 2 demonstrates that TF-IDF is a better metric than other representativeness metrics in both static and dynamic data selection settings. Also, the round-trip BLEU score can be better at measuring the simplicity of sentences than LM-gen. Last, by comparing the Moore-Lewis method ("LM-diff") with our curriculum strategy ("LM-in+LM-gen"), we can see that our method outperform Moore-Lewis method in 3 out of 4 settings.

4.4 Analysis

Next, we investigate how noise in the back-translated data impacts the model performance, how many sentences we should select, and if our weighting methods assign weights appropriately.

Effect of Back-Translation Quality. We try to generate the back-translated data using sampling, greedy search and beam search for iterative back-translation and the results are listed in Table 3. We find that the sampling method significantly degrades the model performance, as it introduces more noise than other approaches, demonstrating that noise can have a negative impact in domain adaptation settings. The conclusion is similar to the findings in low-resource settings Edunov et al. (2018). In addition, we find that our weighting strategies are more beneficial in noisy settings.

Effect of the Percentage p. We test how many sentences should be selected at each epoch for our curriculum strategies. As shown in Figure 3, selecting 30% of the monolingual sentences achieves the best performance in general. Selecting fewer samples can discard valuable information whereas choosing more instances can introduce more noise.

	Back-Translated Sentence	Weight	BLEU
Source	- wenn der Viehhalter seinen Betrieb einem Nachfolger bis zum dritten Verwandtschaftsgrad übergibt ;	-	-
Reference	- when the farmer gives over his farm to his family successor up to the third degree of relationship ,	-	-
Ite-5K	- if the livestock farmer hands over his holding to a successor up to the third degree of kinship ;	0.550	0.353
Ite-10K	- when the livestock farmer passes his holding to a successor up to the third degree of kinship ;	0.572	0.383
Ite-15K	- when the livestock farmer gives his holding to a successor up to the third degree of kinship ;	0.585	0.402
Source	folgerichtig sollte dies auch auf Antisubventionsuntersuchungen zutreffen .	-	-
Reference	the same principles should logically apply to anti - subsidy investigations .	-	-
Ite-5K	this should also be followed up by anti - subsidy investigations .	0.389	0.331
Ite-10K	it should also be folly to apply to anti - subsidy investigations .	0.403	0.486
Ite-15K	it should also be folly true to apply to anti - subsidy investigations .	0.397	0.447

Table 4: Examples of our weighting strategy (Enc). We use our model (Curri+Enc) at the 5K-, 10K-, 15K-th iterative back-translation step to weight sentences. The assigned weights correlate well with the BLEU scores.

R-BLEU	**TF-IDF**	
	High ($\approx$ 1)	**Low** ($\approx$ 0)
High ($\approx$ 1)	Article 20	(2005 / 686 / EC)
Low ($\approx$ 0)	any Contracting Party may request that a meeting be held .	MS Danuta HÜBNER

Table 5: Example full sentences with different TF-IDF and R-BLEU scores. R-BLEU correlates with the lengths while TF-IDF measures the domain distance.

	train	**dev**	**test**	**mono**
low	WMT en-de (100K)	test2013 (3K) LAW (2K) MED (2K)	test2014 (3K) LAW (2K) MED (2K)	CC (1M) LAW (25K) MED (20K)
high	WMT en-de (4.5M)	test2013 (3K)	test2014 (3K)	CC (10M)
high	WMT en-lt (2M)	dev2019 (2K)	test2019 (1K)	News lt (5M) + CC en (5M)

Table 6: Sources and numbers of sentences of the datasets in both low- and high- resource settings. "CC" refers to the CommonCrawl corpus.

Weighting Examples. We use our model (Curri+Enc) to back-translate some sentences from the monolingual corpus and Table 4 shows the weights our models assign at different training stages. In this example, the assigned weights correlate well with the BLEU scores, demonstrating our methods can perform weighting appropriately in some cases.

4.5 Characteristics of the Selected Data

In this part, we investigate certain characteristics of the selected samples.

Lengths. Figure 4 shows the average lengths of the selected sentences in each bucket. We can see that 1) both LM-in and BERT favor long sentences, with one possible explanation being that those sentences are more likely to contain in-domain words; 2) TF-IDF does not share this feature, likely due to the IDF term; 3) sentences with high R-BLEU scores are generally short, likely because NMT models are bad at translating long sentences.

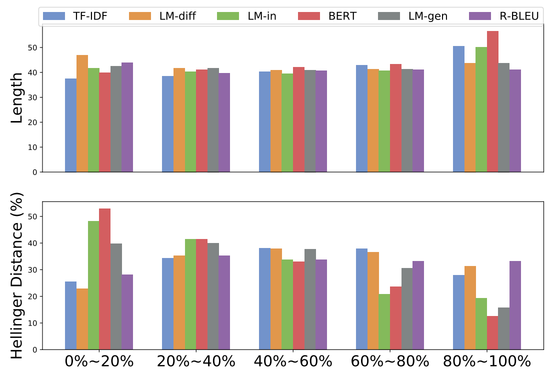

Figure 4: Length and Hellinger distance of sentences in each bucket selected with different metrics.

Unigram Distribution Distance. We also compute the unigram distribution distance using the Hellinger distance. Concretely, we compute the unigram distribution P and Q for both the selected data and the test set, and calculate

$$\frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{V}(\sqrt{p_i}-\sqrt{q_i})^2},$$

where V is the size of the vocabulary. The larger the Hellinger distance is, the more dissimilar the two distributions are. Figure 4 shows that both TF-IDF and BERT match the test distribution well. Also, LM-in performs better than LM-diff, which confirms our hypothesis that the data selected by the Moore-Lewis method cannot adequately represent the target distribution.

Diversity Among Selected Data at Each Epoch. As our curriculum strategies dynamically select different subsets of data, here we examine how many new sentences are actually introduced at each epoch. We find that starting from the second epoch,

Method	WMT-low						WMT-high			
	News		LAW		MED		News		News	
	de-en	en-de	de-en	en-de	de-en	en-de	de-en	en-de	lt-en	en-lt
Baseline										
Base	8.60	6.37	5.51	4.76	6.03	5.19	32.43	27.34	16.24	11.20
Ite-Back	15.80	12.18	20.27	12.41	29.64	21.90	33.02	27.82	19.44	12.41
Best Selection										
Select	15.44	12.09	21.19*	12.70	30.84*	21.97	32.89	27.97	19.52	12.20
Best Curriculum										
Curri	**16.45***	12.61*	21.53*	12.97*	**31.22***	21.71	33.34*	**28.12***	19.82*	12.48
Weighting										
Enc	16.03	12.59*	20.24	12.55	29.95*	**22.18***	32.80	28.03	19.64	12.46
Agree	15.80	12.55*	20.76*	12.85*	29.96*	21.69	32.80	28.00	19.53	**12.66**
Best Curriculum + Weighting										
Curri+Enc	16.24	**12.70**	21.30	12.99	30.82	21.56	33.21	27.97	**20.05**	12.50
Curri+Enc-Diff	16.13	12.65	21.80*	13.18	30.73	21.58	33.15	28.02	19.51	12.39
Curri+Agree	16.23	12.40	21.83*	13.13	30.78	21.66	33.10	27.99	19.73	12.48
Curri+Agree-Diff	16.20	12.61	**22.06***	**13.28***	30.75	21.30	**33.45**	27.91	19.48	12.42

Table 7: Translation accuracy (BLEU) in low-resource and high-resource scenarios. The first and second row list the source and target domains. The third row lists the translation directions. The highest scores are in **bold** and $*$ indicates statistical significance compared with the best baseline ($p < 0.05$).

12.5%, 10.4%, 12.5%, 18.3%, 21.5% of the selected sentences will be replaced at each epoch, and 52.5% of the monolingual sentences will be selected at least once in total.

Examples. Table 5 shows examples of the selected sentences. Sentences with both high TF-IDF and R-BLEU scores are typically short and match the target distribution well. Sentences with high TF-IDF but low R-BLEU scores can be long and contain some out-of-vocabulary words, while sentences with low TF-IDF but high R-BLEU scores are generally short and frequently include digits and single characters. Most of the sentences with both low TF-IDF and R-BLEU scores are extremely noisy and can be safely discarded.

5 Experiments on Low-Resource and High-Resource Scenarios

Next, we conduct experiments in both low- and high-resource scenarios over two language pairs: Lithuanian-English and German-English.

5.1 Setup

Data statistics are shown in Table 6. When the target distribution is the news domain, we train the in-domain LMs with 500K sentences from the news monolingual data. The other settings (including hyperparameters) are the same as before.

5.2 Results

The results are reported in Table 7. We find that LM-in and LM-gen is the best metric combination for curriculum strategies when the target distribution is the news domain. TF-IDF and R-BLEU as the representativeness and simplicity metrics are the best in all other settings.

Low-Resource Settings. In low-resource settings, iterative back-translation can improve the baseline model by a large margin, and our curriculum strategies can still outperform the strong baseline by 1.3 BLEU points. Weighting methods also generally help and in the best case scenario, our method can improve iterative back-translation by 1.8 BLEU points.

High-Resource Settings. In high-resource settings, our curriculum strategies improve the iterative back-translation baseline by up to 0.3 BLEU points. Data weighting strategies do not always help, probably because in high-resource settings the back-translated data is already of high quality. In the best case scenario, our method outperforms iterative back-translation by 0.6 BLEU points.

6 Related Work

Back-translation (Sennrich et al., 2016a) has proven to be effective and several extensions of it have been proposed (He et al., 2016; Cheng et al., 2016; Zhang and Zong, 2016; Xia et al., 2019), among which iterative back-translation methods (Cotterell and Kreutzer, 2018; Hoang et al., 2018; Niu et al., 2018; Zheng et al., 2020) have demonstrated strong empirical performance.

For domain adaptation, Moore and Lewis (2010) and Kirchhoff and Bilmes (2014) use language model cross entropy differences and TF-IDF to select data that are similar to in-domain text respectively. van der Wees et al. (2017) propose dynamic data selection strategies for machine translation models, and Zhang et al. (2019) extend the idea to curriculum strategies. As for filtering noisy sentences, Junczys-Dowmunt (2018) propose to utilize the agreement between forward and backward translation models and Wang et al. (2019a) propose uncertainty-based confidence estimation to improve back-translation. Wang et al. (2019b) compose dynamic domain-data selection with dynamic clean-data selection. Our methods generalize previous data selection strategies and our primary focus is to improve iterative back-translation, but our work could be extended to also include training-time dynamic data selection approaches such as the technique of Wang et al. (2020).

7 Conclusion

In this paper, we provide a novel insight into a widely-used data selection method (Moore and Lewis, 2010) and generalize it to a curriculum strategy for iterative back-translation. We also propose data weighting methods to down-weight examples of poor quality. Extensive experiments are performed to evaluate the performance of our methods; analyses reveal the selected samples can represent the target domain well and our weighting strategies benefit noisy settings the most.

Acknowledgements

The authors are grateful to the anonymous reviewers for their constructive comments, and to Xinyi Wang, Chunting Zhou for their valuable feedback. We also thank Amazon for providing GPU credits.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ryan Cotterell and Julia Kreutzer. 2018. Explaining and generalizing back-translation through wake-sleep. *arXiv*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Erick Fonseca, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *Conference on Machine Translation (WMT)*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Conference on Machine Translation (WMT)*.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Workshop on Neural Generation and Translation (WNGT)*.

Qi Hou, Shujian Huang, Tianhao Ning, Xinyu Dai, and Jiajun Chen. 2019. Nju submissions for the wmt19 quality estimation shared task. In *Conference on Machine Translation (WMT)*.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Workshop on Asian Translation (WAT)*.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Conference on Machine Translation (WMT)*.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M Amin Farajian, António V Lopes, and André FT Martins. 2019. Unbabel's participation in the wmt19 translation quality estimation shared task. In *Conference on Machine Translation (WMT)*.

Hyun Kim and Jong-Hyeok Lee. 2016. A recurrent neural networks approach for estimating the quality of machine translation output. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Katrin Kirchhoff and Jeff Bilmes. 2014. Submodularity for data selection in machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

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 *Meetings of the North American Chapter of the Association for Computational Linguistics (NAACL) Demo Track*.

Xing Niu, Michael Denkowski, and Marine Carpuat. 2018. Bi-directional neural machine translation with synthetic parallel data. *Workshop on Neural Generation and Translation (WNGT)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Lucía Santamaría and Amittai Axelrod. 2019. Data selection with cluster-based language difference models and cynical selection. In *International Conference on Spoken Language Translation (IWSLT)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kashif Shah, Varvara Logacheva, Gustavo Paetzold, Frederic Blain, Daniel Beck, Fethi Bougares, and Lucia Specia. 2015. SHEF-NN: Translation quality estimation with neural networks. In *Workshop on Statistical Machine Translation (WMT)*.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with QuEst++. In *Annual Meeting of the Association for Computational Linguistics (ACL) Demo Track*.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Language Resources and Evaluation Conference (LREC)*.

Marco Turchi, Antonios Anastasopoulos, José GC de Souza, and Matteo Negri. 2014. Adaptive quality estimation for machine translation. In *Annual Meeting of the Association for Computational Linguistics (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 *Conference on Neural Information Processing Systems (NeurIPS)*.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019a. Improving back-translation with uncertainty-based confidence estimation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019b. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xin-Yu Dai, and Jia jun Chen. 2020. Mirror-generative neural machine translation. In *International Conference on Learning Representations (ICLR)*.

Junpei Zhou, Zhisong Zhang, and Zecong Hu. 2019. Source: Source-conditional elmo-style model for machine translation quality estimation. In *Conference on Machine Translation (WMT)*.

A Implementation Details

- We use one 11G NVIDIA GTX 1080 GPUs for each experiment.
- The average training time are: about 30 hours for the baseline models and 40 hours for our models.
- The number of model parameters is 156.81M.
- We use BLEU (Papineni et al., 2002) to evaluate the performance of our models,[2] and *compare-mt* (Neubig et al., 2019) to help with the analysis.[3]
- We manually tune the hyperparameters c_0 in $[0, 0.1, 0.2]$ and T in $[5, 10, 20]$ in Equation 3, and also the percentage of the selected sentences p in each epoch in $[10\%, 20\%, 30\%, 40\%, 50\%]$. We first set c_0 to 0.1, T to 10 and search for the best p, then search for the best T, and finally for c_0, which takes 11 trials in total.
- We follow the instructions on the WMT website to pre-process the data.[4]
- The datasets we use can be downloaded from the WMT website.[5]

[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

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

[4] http://data.statmt.org/wmt17/translation-task/preprocessed/de-en/prepare.sh

[5] http://www.statmt.org/wmt14