Improving Low-Resource Neural Machine Translation with Filtered Pseudo-parallel Corpus

Aizhan Imankulova and **Takayuki Sato** and **Mamoru Komachi**
Tokyo Metropolitan University
{imankulova-aizhan, sato-takayuki}@ed.tmu.ac.jp,
komachi@tmu.ac.jp

Abstract

Large-scale parallel corpora are indispensable to train highly accurate machine translators. However, manually constructed large-scale parallel corpora are not freely available in many language pairs. In previous studies, training data have been expanded using a pseudo-parallel corpus obtained using machine translation of the monolingual corpus in the target language. However, in low-resource language pairs in which only low-accuracy machine translation systems can be used, translation quality is reduces when a pseudo-parallel corpus is used naively. To improve machine translation performance with low-resource language pairs, we propose a method to expand the training data effectively via filtering the pseudo-parallel corpus using a quality estimation based on back-translation. As a result of experiments with three language pairs using small, medium, and large parallel corpora, language pairs with fewer training data filtered out more sentence pairs and improved BLEU scores more significantly.

1 Introduction

A large-scale parallel corpus is an essential resource for training statistical machine translation (SMT) and neural machine translation (NMT) systems. Creating a high-quality large-scale parallel corpus requires time, money and professionals to translate a large amount of texts. As a result, many of the existing large-scale parallel corpora are limited to specific languages and domains. In contrast, large monolingual corpora are easier to obtain.

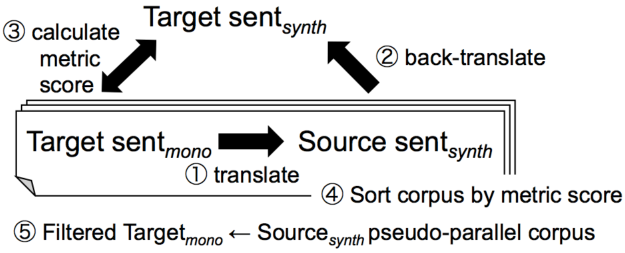

Figure 1: Creating and filtering a pseudo-parallel corpus using back-translation.

Various approaches have been proposed to create a pseudo-parallel corpus from a monolingual corpus. For example, Zhang et al. (2016) proposed a method to generate a pseudo-parallel corpus based on a monolingual corpus of the source language and its automatic translation. Sennrich et al. (2016) obtained substantial improvements by automatically translating a monolingual corpus of the target language into the source language, which they refer to as synthetic, and treating the obtained pseudo-parallel corpus as additional training data. They used monolingual data of the target language to learn the language model more effectively. However, they experimented on language pairs where relatively large-scale parallel corpora are available. Thus, they did not need to fully exploit the training corpus nor care about the quality of the pseudo-parallel corpus.

Therefore, we propose a method to create a pseudo-parallel corpus by back-translating and filtering a monolingual corpus in the target language for low-resource language pairs. If the target sentence and its back-translation are similar, we assume that the synthetic source sentence is appropriate regarding its monolingual target sentence and can be included into the filtered pseudo-parallel corpus. The quality of the pseudo-parallel corpus is especially important because low-quality

Proceedings of the 4th Workshop on Asian Translation, pages 70–78,
Taipei, Taiwan, November 27, 2017. ©2017 AFNLP

parallel sentences will degrade NMT performance more than SMT. Our motivation is to filter out low-quality synthetic sentences that might be included in such a pseudo-parallel corpus to obtain a high-quality pseudo-parallel corpus for low-resource language pairs. To the best of our knowledge, this is the first attempt to (1) filter a pseudo-parallel corpus using back-translation and (2) bootstrap NMT.

The main contributions of our research are as follows:

- We filter a pseudo-parallel corpus using sentence-level similarity metric, in our case sentence-level BLEU (Lin and Och, 2004a,b), and obtain a trainable high-quality pseudo-parallel corpus.
- We show that the proposed filtering method is useful for low-resource language pairs, although bootstrapping does not outperform the proposed filtering method significantly.
- We will release the obtained filtered pseudo-parallel corpora[1].

In this study, we used Japanese↔Russian as low-resource language pairs, French→Malagasy as medium-resource language pairs and German→English as high-resource language pairs. We show that a previous state-of-the-art method (Sennrich et al., 2016) is effective for high-resource language pairs; however, in the case of low-resource language pairs, it is more effective to use a filtered pseudo-parallel corpus as additional training data.

The remainder of this paper is organized as follows: Section 2 discusses previous studies related to improving low-resource machine translation systems; Section 3 outlines the proposed method for filtering a pseudo-parallel corpus and bootstrapping NMT; Sections 4 and 5 evaluate the proposed model; and Section 6 discusses the results. Conclusions and suggestions for future work are presented in Section 7.

2 Related Work

To address the data sparsity problem, there are many methods that use source language monolingual data to improve translation quality (Ueffing et al., 2007; Shwenk, 2008; Bertoldi and Federico, 2009; Hsieh et al., 2013; Zhang et al., 2016). Specifically, Bertoldi and Federico (2009) addressed the problem of domain adaptation by training a translation model from a generated pseudo-parallel corpus created from a monolingual in-domain corpus. Hsieh et al. (2013) create a pseudo-parallel corpus from patterns learned from source and monolingual target in-domain corpora for cross-domain adaptation. They manually conducted filtration of "relatively more accurate" translated sentences and used them to revise the language model. Similarly, we use a pseudo-parallel corpus created by translating a monolingual corpus from the target language rather than the source language; however we apply automatic filtering to the obtained pseudo-parallel corpus.

Data filtering is often used in domain adaptation (Moore and Lewis, 2010; Axelrod et al., 2011) and phrase-based SMT systems. Sentences are extracted from large corpora to optimize the language model and the translation model (Wang et al., 2014; Yıldız et al., 2014). The work most closely related to our work is Yıldız et al. (2014), who build a quality estimator to obtain high-quality parallel sentence pairs and achieve better translation performance and reduce time-complexity with a small high-quality corpus. This method filters data by calculating similarity between source and target sentences. In our work, we calculate similarity between monolingual and synthetic target sentences.

Recently, van der Wees et al. (2017) performed dynamic data selection during training an NMT. To sort and filter the training data, they used language models from the source and target sides of in-domain and out-of-domain data to calculate cross-entropy scores. However, we employ back-translation to filter data considering its meaning.

He et al. (2016) present a dual learning approach. They simultaneously train two models through a reinforcement learning process. They use monolingual data of both source and target languages and generate informative feedback signals to train the translation models. While the dual learning approach is shown to alleviate the issue of noisy data by increasing coverage, we are attempting to remove the noisy data. In addition, they assume a high-recourse language pair to cold start the reinforcement learning process, while we target low-resource language pairs wherein high-quality seed NMT models are difficult to obtain.

[1] https://github.com/aizhanti/filtered-pseudo-parallel-corpora

3 Improving Low-resource Neural Machine Translation (NMT) with Filtered Pseudo-parallel Corpus

In this paper, we propose a method of filtering a pseudo-parallel corpus used as additional training data by back-translating a monolingual corpus for low-resource language pairs. Then, we attempt to bootstrap an NMT model by iterating the filtering process until convergence.

3.1 Filtering

As shown in Figure 1, the proposed method has following steps:

1. Translate monolingual target sentences (Target sent$_{mono}$) using a model trained on parallel corpus in target→source direction to produce synthetic source sentences (Source sent$_{synth}$). Here, we obtain an *"Unfiltered"* pseudo-parallel corpus as additional data without a filtration, similar to Sennrich et al. (2016).
2. Back-translate the synthetic source sentences using a model trained on parallel corpus in source→target direction to obtain a synthetic target sentences (Target sent$_{synth}$).
3. Calculate sentence-level similarity metric scores using the monolingual target sentences as reference and the synthetic target sentences as candidates.
4. Sort the monolingual target sentences and the corresponding synthetic source sentences by a descending order of sentence-level similarity metric scores and filter out sentences with low scores. The threshold is determined by the translation quality on the development set.
5. Use the filtered synthetic source sentences as the source side and the monolingual target sentences as the target side of the pseudo-parallel corpus; this is referred to as a *Filtered* pseudo-parallel corpus as additional data.

3.2 Bootstrapping

Bootstrapping involves the following steps:

1. *"Bootstrap 1"*: we use a pseudo-parallel corpus created using the *"Parallel"* model as additional data to train the seed NMT systems.
2. *"Bootstrap 2"*: we select the best model on the development set from *"Bootstrap 1"* and train its target→source model. Here, we use target sentences from the pseudo-parallel corpus that have been filtered out in the previous iteration to train the best model. If there is no improvement over the previous iteration, terminate the bootstrapping process and return to the *Filtered* pseudo-parallel corpus and the translation model as output. Repeat.

Corpus	Ru↔Ja	Fr→Mg	De→En
Parallel	10,231	106,406	4,535,522
Dev	500	1,000	3,000
Test	500	1,000	3,003
Mono target	75k↔167k	105,570	4,208,439

Table 1: Data statistics.

Even if the monolingual target sentences remain the same, the synthetic source sentences are refreshed at each iteration. In other words, the translation quality of both the *"Unfiltered"* and *"Filtered"* pseudo-parallel corpus will be improved via the bootstrapping process until the termination criterion is met.

4 Experiments Using a Filtered Pseudo-parallel Corpus

4.1 Settings

We used the OpenNMT toolkit[2] (Klein et al., 2017) to train all translation models. For the Russian↔Japanese and French→Malagasy experiments, we used the following parameters: the number of recurrent layers of the encoder and decoder was 1, BiLSTM with concatenation, maximum batch size was 32, and the optimization method was Adadelta. For the German→English experiments, OpenNMT default settings were used. The vocabulary size in all experiments was 50,000.

We tokenized and truecased French, English, German, and Russian sentences using Moses' scripts. For Japanese sentences, we used MeCab 0.996 with the IPAdic dictionary[3] for word segmentation. We eliminated duplicated sentences and sentences with more than 50 words for all languages. We report BLEU scores (Papineni et al., 2002) to compare translation results. We used the Travatar toolkit (Neubig, 2013) to calculate the significance of differences between systems using bootstrap resampling ($p < 0.05$).

[2] http://opennmt.net/OpenNMT/

[3] http://taku910.github.io/mecab

Threshold	Size	Dev	Test
Baselines			
Parallel Ja-Ru	10,231	10.13	9.53
Parallel Ru-Ja	10,231	17.47	**18.71**
Unfiltered	170,991	16.86	17.05
Filtered			
sent-LM ≥ 0.1	168,572	18.65	18.01
sent-LM ≥ 0.2	167,340	17.42	16.65
sent-LM ≥ 0.3	165,166	18.42	16.85
sent-LM ≥ 0.4	160,635	**18.69**	16.23
sent-LM ≥ 0.5	150,974	17.82	17.28
sent-LM ≥ 0.6	131,402	17.37	16.86
sent-LM ≥ 0.7	95,573	17.69	17.54
sent-LM ≥ 0.8	40,774	17.56	16.95
sent-LM ≥ 0.9	11,542	18.13	17.22
sent-LM = 1.0	10,232	18.38	16.93

(a) Bootstrap 1.

Threshold	Size	Dev	Test
Baselines			
B1 Parallel Ja-Ru	160,635	9.05	8.32
B1 Parallel Ru-Ja	160,635	**18.69**	16.23
B1 Unfiltered	170,991	17.03	17.75
Filtered			
sent-LM ≥ 0.1	161,261	16.92	17.63
sent-LM ≥ 0.2	160,866	17.75	16.58
sent-LM ≥ 0.3	160,704	18.29	18.33
sent-LM ≥ 0.4	160,654	18.64	17.37
sent-LM ≥ 0.5	160,640	18.29	17.63

(b) Bootstrap 2.

Table 2: Russian→Japanese translation BLEU scores. Sorting was performed using sent-LM score.

Threshold	Size	Dev	Test
Baselines			
Parallel Ja-Ru	10,231	10.13	9.53
Parallel Ru-Ja	10,231	17.47	18.71
Unfiltered	170,991	16.86	17.05
Filtered			
sent-BLEU ≥ 0.1	26,826	19.86⋆†	19.80⋆†
sent-BLEU ≥ 0.2	24,794	20.29⋆†	19.53⋆†
sent-BLEU ≥ 0.3	19,444	**20.63**⋆†	19.69⋆†
sent-BLEU ≥ 0.4	15,438	20.34⋆†	**20.05**⋆†
sent-BLEU ≥ 0.5	13,101	20.03⋆†	19.35†
sent-BLEU ≥ 0.6	11,904	18.89	19.52†
sent-BLEU ≥ 0.7	11,244	18.79	18.81†
sent-BLEU ≥ 0.8	10,976	18.19	19.21†
sent-BLEU ≥ 0.9	10,867	18.42	17.30
sent-BLEU = 1.0	10,865	18.40	18.45

Table 3: Russian→Japanese translation BLEU scores. Sorting was performed using sent-BLEU score (Bootstrap 1). There is a significant difference: ⋆: against *"Parallel"* baseline; †: against *"Unfiltered"* baseline.

4.2 Dataset

The parallel corpora for low-resource Russian↔Japanese[4] and for medium-resource French→Malagasy[5] experiments were downloaded from OPUS. For the medium-resource French-Malagasy language pair, we used the GlobalVoices corpus, which differs from the Tatoeba corpus used in the previous experiments. Note that the GlobalVoices corpus has more available parallel data (106,406 sentence pairs compared to 10,231).

We split the Tatoeba parallel corpus for the Russian↔Japanese experiments as follows: training set, 10,231 sentences; development set, 500 sentences; and test set, 500 sentences. In addition, to perform Japanese→Russian→Japanese translation for the Russian to Japanese experiment, we sampled an additional 167,600 Japanese monolingual sentences from Tatoeba. We also sampled 75,401 Russian monolingual sentences from Tatoeba for Japanese→Russian translation to facilitate Russian→Japanese→Russian translation.

We performed experiments for the language pair French→Malagasy language pairs using the data from the GlobalVoices corpus. Parallel data were split as follows: training set, 106,406 sentences; development set, 1,000 sentences; and test set, 1,000 sentences. Note that 105,570 Malagasy monolingual sentences from GlobalVoices were used to create a French→Malagasy pseudo-parallel corpus.

For the German→English experiments, we downloaded pre-trained German↔English models and 4,535,522 parallel sentences provided by OpenNMT[6] and used the OpenNMT settings to preprocess all data. We downloaded 4,208,439 German→English sentences from automatically back-translated monolingual data[7] and translated the synthetic German side back to English using

[4] http://opus.lingfil.uu.se/Tatoeba.php

[5] http://opus.lingfil.uu.se/GlobalVoices.php

[6] http://opennmt.net/Models/

[7] http://data.statmt.org/rsennrich/wmt16_backtranslations/de-en/

Threshold	Size	Dev	Test
Baselines			
B1 Parallel Ja-Ru	19,444	12.13	9.78
B1 Parallel Ru-Ja	19,444	20.63†	19.69†
B1 Unfiltered	170,991	18.06	16.85
Filtered			
sent-BLEU $\geq$ 0.1	40,567	21.03†	21.01†
sent-BLEU $\geq$ 0.2	37,531	**21.48**†	19.20†
sent-BLEU $\geq$ 0.3	29,533	21.06†	20.69†
sent-BLEU $\geq$ 0.4	24,290	21.16†	21.08†
sent-BLEU $\geq$ 0.5	21,742	20.58†	**21.57**⋆†
sent-BLEU $\geq$ 0.6	20,478	19.93†	20.80†
sent-BLEU $\geq$ 0.7	19,920	20.46†	20.48†
sent-BLEU $\geq$ 0.8	19,726	20.78†	20.60†
sent-BLEU $\geq$ 0.9	19,626	20.38†	21.54⋆†
sent-BLEU = 1.0	19,623	21.23†	21.17⋆†

Table 4: Russian→Japanese translation BLEU scores. Sorting was performed using sent-BLEU score (Bootstrap 2).

Threshold	Size	Dev	Test
Baselines			
B2 Parallel Ja-Ru	37,531	12.35	11.78
B2 Parallel Ru-Ja	37,531	**21.48**†	19.20†
B2 Unfiltered	170,991	18.96	17.20
Filtered			
sent-BLEU $\geq$ 0.1	53,478	21.34†	19.10†
sent-BLEU $\geq$ 0.2	49,833	20.61†	19.99†
sent-BLEU $\geq$ 0.3	43,470	21.32†	20.59†
sent-BLEU $\geq$ 0.4	40,147	20.75†	20.16†
sent-BLEU $\geq$ 0.5	38,687	20.40†	18.65
sent-BLEU $\geq$ 0.6	38,043	20.03	**21.02**⋆†
sent-BLEU $\geq$ 0.7	37,758	20.17†	20.23†
sent-BLEU $\geq$ 0.8	37,639	20.33†	20.61†
sent-BLEU $\geq$ 0.9	37,600	19.75	19.80†
sent-BLEU = 1.0	37,598	20.83†	20.62†

Table 5: Russian→Japanese translation BLEU scores. Sorting was performed using sent-BLEU score (Bootstrap 3).

the pre-trained German→English model to filter this pseudo-parallel corpus. We used newtest2013 (3,000 sentence pairs) as a development set and newtest2014 (3,003 sentence pairs) as a test set. Table 1 shows the data statistics.

4.3 Baselines

Sennrich et al. (2016) obtained additional training data by automatically translating monolingual target sentences into the source language using their *"Parallel"* baseline system. Our process differs from theirs in that we construct *"Parallel"* baseline machine translation systems in both directions using an available parallel corpus to obtain a filtered pseudo-parallel corpus.

Our baseline systems were as follows: 1) *"Parallel"* systems that trained on a parallel corpus in both directions, which were used to create a pseudo-parallel corpus; or *"B{1,2} Parallel"* in case of bootstrapping 2) *"Unfiltered"* system, which was trained on a concatenated parallel corpus with all pseudo-parallel corpora without filtration; or *"B{1,2} Unfiltered"* in case of bootstrapping.

4.4 Sentence-level similarity metric

We used sentence-level BLEU (sent-BLEU) as a sentence-level similarity metric. The sent-BLEU scores were calculated using mteval-sentence of the mteval toolkit[8]. In Russian→Japanese experiments, we compared the sent-BLEU scores, which require back-translation of the target monolingual data for the proposed filtration method, with a language model (sent-LM) that performs filtration by scoring only synthetic source sentences. We used the KenLM Language Model Toolkit[9] to build a 5-gram language model from 23,239,280 sentences from the Russian side of the Russian-English UN corpus (Ziemski et al., 2016).[10]. We also applied Kneser-Ney smoothing. To extract the scores, we normalized the language model log probability of the sentence to be between [0, 1] as in sent-BLEU using a feature scaling method.

Translation performance increases as the number of parallel sentences increases (Koehn, 2002). For a pseudo-parallel corpus, however, translation performance does not necessarily increase with the number of sentences. To determine the effects of the quantity and quality of the pseudo-parallel corpus in machine translation, we set thresholds with increment steps of 0.1. Thus, pseudo-parallel sentences included as additional data have sentence-level similarity scores greater or equal to some threshold (e.g., sentence-level BLEU$\geq$ 0.1,..., sentence-level BLEU$\geq$ 0.9, ...). Sentences scored and filtered by sentence-level similarity were used to train *"Filtered"* models. For example, sentences with sentence-level sim-

[8] https://github.com/odashi/mteval

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

[10] https://conferences.unite.un.org/UNCorpus/en/DownloadOverview

Threshold	Size	Dev	Test
Baselines			
Parallel Ru-Ja	10,231	17.47	18.71
Parallel Ja-Ru	10,231	10.13	9.53
Unfiltered	85,632	10.40	9.01
Filtered			
sent-BLEU $\geq$ 0.1	12,686	12.86⋆†	12.81⋆†
sent-BLEU $\geq$ 0.2	12,613	12.82⋆†	13.60⋆†
sent-BLEU $\geq$ 0.3	12,325	**14.08**⋆†	13.34⋆†
sent-BLEU $\geq$ 0.4	11,860	13.14⋆†	**14.08**⋆†
sent-BLEU $\geq$ 0.5	11,462	11.95⋆†	13.86⋆†
sent-BLEU $\geq$ 0.6	11,114	11.92⋆†	11.50⋆†
sent-BLEU $\geq$ 0.7	10,965	12.34⋆†	12.73⋆†
sent-BLEU $\geq$ 0.8	10,903	12.30⋆†	11.81⋆†
sent-BLEU = 1.0	10,880	11.69⋆	11.52⋆†

Table 6: Japanese→Russian translation BLEU scores. Sorting was performed using sent-BLEU score.

Threshold	Size	Dev	Test
Baselines			
Parallel Mg-Fr	106,406	13.29	12.74
Parallel Fr-Mg	106,406	16.79	15.15
Unfiltered	211,976	16.39	14.80
Filtered			
sent-BLEU $\geq$ 0.1	152,578	**17.31**	**16.27**⋆†
sent-BLEU $\geq$ 0.2	135,179	17.08	15.33
sent-BLEU $\geq$ 0.3	121,376	17.11	15.00
sent-BLEU $\geq$ 0.4	114,391	16.62	15.81
sent-BLEU $\geq$ 0.5	110,944	16.65	14.84
sent-BLEU $\geq$ 0.6	109,186	16.38	14.05
sent-BLEU $\geq$ 0.7	108,252	16.48	15.19
sent-BLEU $\geq$ 0.8	107,801	16.29	14.53
sent-BLEU $\geq$ 0.9	107,537	16.42	15.24
sent-BLEU = 1.0	107,515	16.38	15.26

Table 7: French→Malagasy translation BLEU scores. Sorting was performed using sent-BLEU score.

ilarity scores (e.g., sent-BLEU) greater than or equal to 0.1 were used to train the *"sent-BLEU $\geq$ 0.1"* model. We trained the NMT system using different thresholds and compared the performance using development and test sets.

5 Results

5.1 Bootstrapping the NMT: Russian→Japanese

For the data shown in Tables 2 and 3, we used the parallel 10,231 sentence pairs (Section 4.2) to train the first *"Parallel"* models in both directions. Then, we used these models to create a pseudo-parallel corpus by translating 160,760 Japanese monolingual sentences (Section 3). A concatenation of parallel and pseudo-parallel sentences was used to train the *"Unfiltered"* model. The results obtained using the *"Unfiltered"* model demonstrate that using all pseudo-parallel data as additional data results in reduced BLEU scores (16.86 BLEU compared to 17.47 BLEU). Generally, these results suggest that unfiltered data contain many incorrect sentence pairs, which leads to reduced machine translation accuracy.

Tables 2a and 3 show the *"Bootstrap 1"* results. Here, the same pseudo-parallel corpus was used as additional data with different filtration scoring metrics. Even though the models trained using data sorted by a language model metric outperformed the baselines on the development set, none of the sent-LM models achieved better results than sent-BLEU. In contrast, using sent-BLEU increased performance even when much less data were used for training. The *"sent-BLEU $\geq$ 0.3"* model outperformed the *"Unfiltered"* model by +3.77 and +2.64 points on the development and test sets, respectively. A sent-LM model resulted in lower BLEU scores compared to sent-BLEU because it assigned high scores to very short but grammatically correct synthetic sentences. For example, a sent-LM assigned a score of 0.94 to the synthetic Russian sentence "д а . *(yes .)*", even though its corresponding monolingual sentence was "歌える 。*(I can sing .)*". In contrast, sent-BLEU assigned this pseudo-parallel sentence a score of 0.00, because the back-translation resulted in "はい 。*(yes .)*". Furthermore, for a sent-LM, the bootstrapping attempt using the best *"sent-LM $\geq$ 0.4"* model of *"Bootstrap 1"* failed according to the results shown in Table 2b. None of the *"Filtered"* models could outperform the *"Bootstrap 1"* and *"Bootstrap 2"* baseline models.

Table 4 shows the *"Bootstrap 2"* results. We used the best model, i.e., *"sent-BLEU $\geq$ 0.3"* from *"Bootstrap 1"* (referred to as *"B1 Parallel"*), to create a pseudo-parallel corpus by translating the filtered out Japanese monolingual sentences (with sent-BLEU < 0.3). The resulting 151,547 pseudo-parallel sentences were added to the 37,531 *"B1 Parallel"* sentences to train the *"B1 Unfiltered"* model. The filtered *"sent-BLEU $\geq$ 0.2"* model

Threshold	Size	Dev	Test
Baselines			
Parallel En-De	4,535,522	19.51	18.55
Parallel De-En	4,535,522	22.33	20.58
Unfiltered	8,743,961	**25.09⋆**	**24.86⋆**
Filtered			
sent-BLEU $\geq$ 0.1	7,681,105	24.84⋆	24.52⋆
sent-BLEU $\geq$ 0.2	7,345,367	24.87⋆	24.13⋆
sent-BLEU $\geq$ 0.3	6,598,845	23.06⋆	22.65⋆
sent-BLEU $\geq$ 0.4	5,808,701	24.13⋆	22.84⋆
sent-BLEU $\geq$ 0.5	5,216,440	23.73⋆	22.28⋆
sent-BLEU $\geq$ 0.6	7,345,367	23.50⋆	21.85⋆
sent-BLEU $\geq$ 0.7	6,598,845	23.07⋆	21.30⋆
sent-BLEU $\geq$ 0.8	5,808,701	22.80⋆	20.90⋆
sent-BLEU $\geq$ 0.9	5,216,440	22.60⋆	20.49
sent-BLEU = 1.0	4,585,655	22.13	20.33

Table 8: German→English translation BLEU scores. Sorting was performed using sent-BLEU score.

was the best model in *"Bootstrap 2"*. This model achieved a 21.48 BLEU score on the development set, thereby outperforming the *"B1 Parallel"* model by +0.85 BLEU points.

The *"Bootstrap 3"* results are shown in Table 5. In the third iteration, no *"Filtered"* models obtained higher scores than the *"B2 Parallel"* model. However, in the Russian→Japanese experiments, all *"Filtered"* models outperformed the *"Unfiltered"* models on the development and test sets in each *"Bootstrap"* step for sentence-level BLEU scoring, demonstrating a maximum improvement of +3.77 BLEU points on the development set and +4.72 BLEU points on the test set.

5.2 Filtering

5.2.1 Japanese→Russian

We examined the effect of the proposed filtering method on Japanese to Russian translations. The results are shown in Table 6. Here, we used a Russian monolingual corpus to create a Japanese→Russian parallel corpus rather than using the Japanese monolingual corpus.

The *"sent-BLEU $\geq$ 0.3"* model outperformed the *"Parallel"* and *"Unfiltered"* models in terms of BLEU scores on the development set by +3.95 and +3.68 points, respectively. All filtered models were significantly better than the unfiltered model, except for *"sent-BLEU = 1.0"*.

Threshold	Ja-Ru	Ru-Ja	Fr-Mg	En-De
sent-BLEU $\geq$ 0.1	3.26%	10.32%	43.73%	74.74%
sent-BLEU $\geq$ 0.2	3.16%	9.06%	27.25%	66.77%
sent-BLEU $\geq$ 0.3	2.78%	5.73%	14.18%	43.09%
sent-BLEU $\geq$ 0.4	2.16%	3.24%	7.56%	30.25%
sent-BLEU $\geq$ 0.5	1.63%	1.79%	4.30%	16.18%
sent-BLEU $\geq$ 0.6	1.17%	1.04%	2.63%	7.88%
sent-BLEU $\geq$ 0.7	0.97%	0.63%	1.75%	3.85%
sent-BLEU $\geq$ 0.8	0.89%	0.46%	1.32%	1.98%
sent-BLEU $\geq$ 0.9	0.89%	0.40%	1.07%	1.25%
sent-BLEU = 1.0	0.86%	0.39%	1.05%	1.19%

Table 9: The percentage of used pseudo-parallel corpora for each language pair.

5.2.2 French→Malagasy

The results are shown in Table 7. In these experiments, we used the Malagasy monolingual corpus comprising 105,570 sentences to create a French-Malagasy pseudo-parallel corpus using the proposed filtering method. The *"sent-BLEU $\geq$ 0.1"* model yielded better results over the baselines of up to +0.92 BLEU points on the development set and +1.47 BLEU points on the test set (statistically significant).

5.2.3 German→English

Table 8 shows the BLEU scores of German→English experiments. None of the filtered models outperformed the *"Unfiltered"* baseline on the development and test sets.

6 Discussion

The results showed that rather than using all additional pseudo-parallel data, the proposed filtering method improved translation performance in nearly all experiments conducted for low-resource language pairs.

The threshold results (Section 4.4) in Tables 2-8 demonstrate that filtered models outperform the baselines with larger margin for low-resource language pairs than high-resource language pair and in the most cases, overfiltering (e.g., sent-BLEU = 1.0) leads to no or negligible improvement over the baselines.

Sennrich et al. (2016) showed that using a pseudo-parallel corpus as additional data greatly improves the performance over the *"Parallel"* baseline. The experiments showed that a better *"Parallel"* system results in the creation of a better pseudo-parallel corpus. This fact is also demonstrated in Table 9, in which the percentages of used pseudo-parallel corpora for each language

Boot	Synthetic Russian sentence	Synthetic Japanese sentence	sent-BLEU
Example 1 - Japanese monolingual sentence: あなた は その ニュース を 聞き ました か 。 **(have you heard the news ?)**			
B1	ты видели эту по-английски ? (did you see this in English ?)	君 は 英語 を 英語 を 見 ました か 。 (have you seen English in English ?)	0.25
B2	Вы получили эту радио ? (did you get this radio ?)	その ニュース を 借り た の です か 。 (did you borrow the news ?)	0.00
B3	Вы получили эту новости ? (did you receive this news ?)	その ニュース を 聞き ました か 。 (have you heard the news ?)	0.77
Example 2 - Japanese monolingual sentence: 僕 は 終電車 に 乗り遅れ た 。 **(I missed the last train .)**			
B1	я опоздал на поезд . (I missed the train .)	私 は 列車 に 遅刻 し た 。 (I was late for the train .)	0.00
B2	я опоздал на поезд . (I missed the train .)	私 は 列車 に 遅れ た 。 (I was late for the train .)	0.00
B3	я опоздал на последний поезд . (I missed the last train .)	私 は 終電車 に 乗り遅れ た 。 (I missed the last train .)	0.80
Example 3 - Japanese monolingual sentence: なぜ 遅刻 し た の です か 。 **(why were you late .)**			
B1	почему ты сделал ? (why did you do it ?)	どうして やっ た の ? (why did it ?)	0.00
B2	почему ты опоздал ? (why are you late ?)	なぜ そんな 遅れ た の ? (why was such a delay?)	0.00
B3	почему ты сделал ? (why did you do it ?)	なぜ そんな こと を し た の です か 。 (why did a such thing ?)	0.53

Table 10: Examples from Russian→Japanese pseudo-parallel corpus used on every bootstrapping step.

pair are shown. The size of the usable pseudo-parallel corpus for low-resource language pairs is very small, which indicates that filtering out very noisy data (e.g., approximately 96%-98% data for Japanese→Russian) results in higher accuracy of the NMT system trained using a filtered pseudo-parallel corpus. The size of very noisy data for a high-resource language pair (e.g. approximately 25% of the data for German→English) is small and does not significantly degrade the accuracy of the NMT system compared to low-resource cases. In other words, the weaker the *"Parallel"* system is the more effective is the proposed filtration method.

Example 1 in the Table 10 shows the steps required to create a better Russian-Japanese pseudo-parallel sentence. As the synthetic Russian sentence from *"Bootstrap 1"* which was significantly incorrect relative to the correct translation of the Japanese monolingual sentences, eventually became a good translation, we can say that the Japanese→Russian and Russian→Japanese models used to create a pseudo-parallel corpus improved with each bootstrapping step. Example 2 in Table 10 shows good translations of the original sentence; however, due to surface mismatching of the synthetic and monolingual target sentences, the sentence-level BLEU scores were 0.00. Nonetheless, with *"Bootstrap 3"*, the Japanese→Russian and Russian→Japanese models produced translations that were the closest to the original sentence. Regarding Example 3, the sentence in *"Bootstrap 2"* was not used to train the best model due to surface mismatching of target sentences despite the fact that it was correctly translated to Russian. As a result, *"Bootstrap 3"* used an incorrect translation of the original sentence.

The experimental results show that bootstrapping over several iterations improves the NMT without significant difference and eventually stops improving over the previous step. We hypothesize that the reason for this is that the *"Parallel"* system used to create a new pseudo-parallel corpus becomes weaker in each iteration.

We used sent-BLEU to calculate the similarity of the synthetic and monolingual target sentences. However, word embedding-based sentence similarity measures, such as those employed by Song and Roth (2015), can be used to further improve the corpus filtering because sentence-level BLEU is sensitive to surface mismatch.

7 Conclusion

The models trained using the filtered pseudo-parallel corpus as additional data showed better translation performance than the baselines for low-resource language pairs. We have also shown that we can further improve translation performance by bootstrapping, although bootstrapping has its limitations. These results suggest that translation accuracy depends on both data size and quality.

Further experimental investigations are required to estimate the limitations of the proposed filtration method. We plan to investigate the other sentence similarity metrics described in Song and Roth (2015), such as average alignment and maximum alignment sentence-level word2vec scores. Sentence-level BLEU calculates the similarity of the synthetic and monolingual target sentences based solely on surface information, whereas word2vec uses a distributed representation of the sentences.

To further our research we plan to improve our filtering method by detecting good and bad synthetic translations using reinforcement learning.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 355–362.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. pages 182–189.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*. pages 820–828.

An-Chang Hsieh, Hen-Hsen Huang, and Hsin-Hsi Chen. 2013. Uses of monolingual in-domain corpora for cross-domain adaptation with hybrid MT approaches. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*. pages 117–122.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1701.02810* .

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation.

Chin-Yew Lin and Franz Josef Och. 2004a. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. pages 605–612.

Chin-Yew Lin and Franz Josef Och. 2004b. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*. pages 501–507.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*. Association for Computational Linguistics, pages 220–224.

Graham Neubig. 2013. Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers. In *the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. pages 91–96.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311–318.

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

Yangqiu Song and Dan Roth. 2015. Unsupervised sparse vector densification for short text similarity. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1275–1280.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. *arXiv preprint arXiv:1708.00712. Accepted at EMNLP2017* .

Longyue Wang, Derek F Wong, Lidia S Chao, Yi Lu, and Junwen Xing. 2014. A systematic comparison of data selection criteria for SMT domain adaptation. *The Scientific World Journal* 2014.

Eray Yıldız, Ahmed Cüneyd Tantuğ, and Banu Diri. 2014. The effect of parallel corpus quality vs size in English-to-Turkish SMT. In *Proceedings of the Sixth International Conference on Web services and Semantic Technology (WeST 2014)*. pages 21–30.

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

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. pages 3530–3534.