Exploiting Out-of-Domain Parallel Data through Multilingual Transfer Learning for Low-Resource Neural Machine Translation

Aizhan Imankulova[†] **Raj Dabre**[‡] **Atsushi Fujita**[‡] **Kenji Imamura**[‡]
[†]Tokyo Metropolitan University
6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
imankulova-aizhan@ed.tmu.ac.jp
[‡]National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{raj.dabre, atsushi.fujita, kenji.imamura}@nict.go.jp

Abstract

This paper proposes a novel multilingual multistage fine-tuning approach for low-resource neural machine translation (NMT), taking a challenging Japanese–Russian pair for benchmarking. Although there are many solutions for low-resource scenarios, such as multilingual NMT and back-translation, we have empirically confirmed their limited success when restricted to in-domain data. We therefore propose to exploit out-of-domain data through transfer learning, by using it to first train a multilingual NMT model followed by multistage fine-tuning on in-domain parallel and back-translated pseudo-parallel data. Our approach, which combines domain adaptation, multilingualism, and back-translation, helps improve the translation quality by more than 3.7 BLEU points, over a strong baseline, for this extremely low-resource scenario.

1 Introduction

Neural machine translation (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) has enabled end-to-end training of a translation system without needing to deal with word alignments, translation rules, and complicated decoding algorithms, which are the characteristics of phrase-based statistical machine translation (PBSMT) (Koehn et al., 2007). Although NMT can be significantly better than PBSMT in resource-rich scenarios, PBSMT performs better in low-resource scenarios (Koehn and Knowles, 2017). Only by exploiting cross-lingual transfer learning techniques (Firat et al., 2016; Zoph et al., 2016; Kocmi and Bojar, 2018), can the NMT performance approach PBSMT performance in low-resource scenarios.

However, such methods usually require an NMT model trained on a resource-rich language pair like French↔English (parent), which is to be fine-tuned for a low-resource language pair like Uzbek↔English (child). On the other hand, multilingual approaches (Johnson et al., 2017) propose to train a single model to translate multiple language pairs. However, these approaches are effective only when the parent target or source language is relatively resource-rich like English (En). Furthermore, the parents and children models should be trained on similar domains; otherwise, one has to take into account an additional problem of domain adaptation (Chu et al., 2017).

In this paper, we work on a linguistically distant and thus challenging language pair Japanese↔Russian (Ja↔Ru) which has only 12k lines of news domain parallel corpus and hence is extremely resource-poor. Furthermore, the amount of indirect in-domain parallel corpora, i.e., Ja↔En and Ru↔En, are also small. As we demonstrate in Section 4, this severely limits the performance of prominent low-resource techniques, such as multilingual modeling, back-translation, and pivot-based PBSMT. To remedy this, we propose a novel multistage fine-tuning method for NMT that combines multilingual modeling (Johnson et al., 2017) and domain adaptation (Chu et al., 2017).

We have addressed two important research questions (RQs) in the context of extremely low-resource machine translation (MT) and our explorations have derived rational contributions (CTs) as follows:

RQ1. What kind of translation quality can we obtain in an extremely low-resource scenario?

CT1. We have made extensive comparisons with multiple architectures and MT paradigms to show how difficult the problem is. We have also explored the utility of back-translation and show that it is ineffective given the poor performance of base MT systems used to generate pseudo-parallel data. Our systematic exploration shows that multilingualism is extremely useful for in-domain translation with very limited corpora (see Section 4). This type of exhaustive exploration has been missing from most existing works.

RQ2. What are the effective ways to exploit out-of-domain data for extremely low-resource in-domain translation?

CT2. Our proposal is to first train a multilingual NMT model on out-of-domain Ja↔En and Ru↔En data, then fine-tune it on in-domain Ja↔En and Ru↔En data, and further fine-tune it on Ja↔Ru data (see Section 5). We show that this stage-wise fine-tuning is crucial for high-quality translation. We then show that the improved NMT models lead to pseudo-parallel data of better quality. This data can then be used to improve the performance even further thereby enabling the generation of better pseudo-parallel data. By iteratively generating pseudo-parallel data and fine-tuning the model on said data, we can achieve the best performance for Japanese↔Russian translation.

To the best of our knowledge, we are the first to perform such an extensive evaluation of extremely low-resource MT problem and propose a novel multilingual multistage fine-tuning approach involving multilingual modeling and domain adaptation to address it.

2 Our Japanese–Russian Setting

In this paper, we deal with Ja↔Ru news translation. This language pair is very challenging because the languages involved have completely different writing system, phonology, morphology, grammar, and syntax. Among various domains, we experimented with translations in the news domain, considering the importance of sharing news between different language speakers. Moreover, news domain is one of the most challenging tasks, due to large presence of out-of-vocabulary (OOV) tokens and long sentences.[1] To establish and evaluate existing methods, we also involved English as the third language. As direct parallel corpora are scarce, involving a language such as English for pivoting is quite common (Utiyama and Isahara, 2007).

Ru	Ja	En	#sent.	Usage	
				test	development
✓	✓	✓	913	600	313
✓	✓		173	-	173
	✓	✓	276	-	276
✓		✓	0	-	-
✓			4	-	-
	✓		287	-	-
		✓	1	-	-
Total			1,654	-	-

Table 1: Manually aligned News Commentary data.

There has been no clean held-out parallel data for Ja↔Ru and Ja↔En news translation. Therefore, we manually compiled development and test sets using News Commentary data[2] as a source. Since the given Ja↔Ru and Ja↔En data share many lines in the Japanese side, we first compiled tri-text data. Then, from each line, corresponding parts across languages were manually identified, and unaligned parts were split off into a new line. Note that we have never merged two or more lines. As a result, we obtained 1,654 lines of data comprising trilingual, bilingual, and monolingual segments (mainly sentences) as summarized in Table 1. Finally, for the sake of comparability, we randomly chose 600 trilingual sentences to create a test set, and concatenated the rest of them and bilingual sentences to form development sets.

Our manually aligned development and test sets are publicly available.[3]

3 Related Work

Koehn and Knowles (2017) showed that NMT is unable to handle low-resource language pairs as opposed to PBSMT. Transfer learning approaches (Firat et al., 2016; Zoph et al., 2016; Kocmi and Bojar, 2018) work well when a large helping parallel corpus is available. This restricts one of the source or the target languages to be English which, in our case, is not possible. Approaches involving bi-directional NMT modeling is shown to drasti-

[1] News domain translation is also the most competitive tasks in WMT indicating its importance.

[2] http://opus.nlpl.eu/News-Commentary-v11.php

[3] https://github.com/aizhanti/JaRuNC

cally improve low-resource translation (Niu et al., 2018). However, like most other, this work focuses on translation from and into English.

Remaining options include (a) unsupervised MT (Artetxe et al., 2018; Lample et al., 2018; Marie and Fujita, 2018), (b) parallel sentence mining from non-parallel or comparable corpora (Utiyama and Isahara, 2003; Tillmann and Xu, 2009), (c) generating pseudo-parallel data (Sennrich et al., 2016), and (d) MT based on pivot languages (Utiyama and Isahara, 2007). The linguistic distance between Japanese and Russian makes it extremely difficult to learn bilingual knowledge, such as bilingual lexicons and bilingual word embeddings. Unsupervised MT is thus not promising yet, due to its heavy reliance on accurate bilingual word embeddings. Neither does parallel sentence mining, due to the difficulty of obtaining accurate bilingual lexicons. Pseudo-parallel data can be used to augment existing parallel corpora for training, and previous work has reported that such data generated by so-called back-translation can substantially improve the quality of NMT. However, this approach requires base MT systems that can generate somewhat accurate translations. It is thus infeasible in our scenario, because we can obtain only a weak system which is the consequence of an extremely low-resource situation. MT based on pivot languages requires large in-domain parallel corpora involving the pivot languages. This technique is thus infeasible, because the in-domain parallel corpora for Ja↔En and Ru↔En pairs are also extremely limited, whereas there are large parallel corpora in other domains. Section 4 empirically confirms the limit of these existing approaches.

Fortunately, there are two useful transfer learning solutions using NMT: (e) multilingual modeling to incorporate multiple language pairs into a single model (Johnson et al., 2017) and (f) domain adaptation to incorporate out-of-domain data (Chu et al., 2017). In this paper, we explore a novel method involving step-wise fine-tuning to combine these two methods. By improving the translation quality in this way, we can also increase the likelihood of pseudo-parallel data being useful to further improve translation quality.

4 Limit of Using only In-domain Data

This section answers our first research question, [RQ1], about the translation quality that we can achieve using existing methods and in-domain parallel and monolingual data. We then use the strongest model to conduct experiments on generating and utilizing back-translated pseudo-parallel data for augmenting NMT. Our intention is to empirically identify the most effective practices as well as recognize the limitations of relying only on in-domain parallel corpora.

Lang.pair	Partition	#sent.	#tokens	#types
Ja↔Ru	train	12,356	341k / 229k	22k / 42k
	development	486	16k / 11k	2.9k / 4.3k
	test	600	22k / 15k	3.5k / 5.6k
Ja↔En	train	47,082	1.27M / 1.01M	48k / 55k
	development	589	21k / 16k	3.5k / 3.8k
	test	600	22k / 17k	3.5k / 3.8k
Ru↔En	train	82,072	1.61M / 1.83M	144k / 74k
	development	313	7.8k / 8.4k	3.2k / 2.3k
	test	600	15k / 17k	5.6k / 3.8k

Table 2: Statistics on our in-domain parallel data.

4.1 Data

To train MT systems among the three languages, i.e., Japanese, Russian, and English, we used all the parallel data provided by Global Voices,[4] more specifically those available at OPUS.[5] Table 2 summarizes the size of train/development/test splits used in our experiments. The number of parallel sentences for Ja↔Ru is 12k, for Ja↔En is 47k, and for Ru↔En is 82k. Note that the three corpora are not mutually exclusive: 9k out of 12k sentences in the Ja↔Ru corpus were also included in the other two parallel corpora, associated with identical English translations. This puts a limit on the positive impact that the helping corpora can have on the translation quality.

Even when one focuses on low-resource language pairs, we often have access to larger quantities of in-domain monolingual data of each language. Such monolingual data are useful to improve quality of MT, for example, as the source of pseudo-parallel data for augmenting training data for NMT (Sennrich et al., 2016) and as the training data for large and smoothed language models for PBSMT (Koehn and Knowles, 2017). Table 3 summarizes the statistics on our monolingual corpora for several domains including the news domain. Note that we removed from the Global Voices monolingual corpora those sentences that are already present in the parallel corpus.

[4] https://globalvoices.org/
[5] http://opus.nlpl.eu/GlobalVoices-v2015.php

Corpus	Ja	Ru	En
Global Voices[5]	26k	24k	842k
Wikinews[6]	37k	243k	-
News Crawl[7]	-	72M	194M
Yomiuri (2007–2011)[8]	19M	-	-
IWSLT[9]	411k	64k	66k
Tatoeba[10]	5k	58k	208k

Table 3: Number of lines in our monolingual data. Whereas the first four are from the news corpora (in-domain), the last two, i.e., "IWSLT" and "Tatoeba," are from other domains.

We tokenized English and Russian sentences using *tokenizer.perl* of `Moses` (Koehn et al., 2007).[11] To tokenize Japanese sentences, we used `MeCab`[12] with the IPA dictionary. After tokenization, we eliminated duplicated sentence pairs and sentences with more than 100 tokens for all the languages.

4.2 MT Methods Examined

We began with evaluating standard MT paradigms, i.e., PBSMT (Koehn et al., 2007) and NMT (Sutskever et al., 2014). As for PBSMT, we also examined two advanced methods: pivot-based translation relying on a helping language (Utiyama and Isahara, 2007) and induction of phrase tables from monolingual data (Marie and Fujita, 2018).

As for NMT, we compared two types of encoder-decoder architectures: attentional RNN-based model (RNMT) (Bahdanau et al., 2015) and the Transformer model (Vaswani et al., 2017). In addition to standard uni-directional modeling, to cope with the low-resource problem, we examined two multi-directional models: bi-directional model (Niu et al., 2018) and multi-to-multi (M2M) model (Johnson et al., 2017).

After identifying the best model, we also examined the usefulness of a data augmentation method based on back-translation (Sennrich et al., 2016).

PBSMT Systems

First, we built a PBSMT system for each of the six translation directions. We obtained phrase tables from parallel corpus using `SyMGIZA++`[13] with the `grow-diag-final` heuristics for word alignment, and `Moses` for phrase pair extraction. Then, we trained a bi-directional MSD (monotone, swap, and discontinuous) lexicalized reordering model. We also trained three 5-gram language models, using `KenLM`[14] on the following monolingual data: (1) the target side of the parallel data, (2) the concatenation of (1) and the monolingual data from Global Voices, and (3) the concatenation of (1) and all monolingual data in the news domain in Table 3.

Subsequently, using English as the pivot language, we examined the following three types of pivot-based PBSMT systems (Utiyama and Isahara, 2007; Cohn and Lapata, 2007) for each of Ja→Ru and Ru→Ja.

Cascade: 2-step decoding using the source-to-English and English-to-target systems.

Synthesize: Obtain a new phrase table from synthetic parallel data generated by translating English side of the target–English training parallel data to the source language with the English-to-source system.

Triangulate: Compile a new phrase table combining those for the source-to-English and English-to-target systems.

Among these three, triangulation is the most computationally expensive method. Although we had filtered the component phrase tables using the statistical significance pruning method (Johnson et al., 2007), triangulation can generate an enormous number of phrase pairs. To reduce the computational cost during decoding and the negative effects of potentially noisy phrase pairs, we retained for each source phrase s only the k-best translations t according to the forward translation probability $\phi(t|s)$ calculated from the conditional probabilities in the component models as defined in Utiyama and Isahara (2007). For each of the retained phrase pairs, we also calculated the backward translation probability, $\phi(s|t)$, and lexical translation probabilities, $\phi_{lex}(t|s)$ and $\phi_{lex}(s|t)$, in the same manner as $\phi(t|s)$.

We also investigated the utility of recent advances in unsupervised MT. Even though we began with a publicly available implementation of

[6] `https://dumps.wikimedia.org/backup-index.html` (20180501)
[7] `http://www.statmt.org/wmt18/translation-task.html`
[8] `https://www.yomiuri.co.jp/database/glossary/`
[9] `http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/`
[10] `http://opus.nlpl.eu/Tatoeba-v2.php`
[11] `https://github.com/moses-smt/mosesdecoder`
[12] `http://taku910.github.io/mecab`, version 0.996.
[13] `https://github.com/emjotde/symgiza-pp`
[14] `https://github.com/kpu/kenlm`

ID	System	Parallel data			Total size of training data	Vocabulary size
		Ja↔Ru	Ja↔En	Ru↔En		
(a1), (b1)	Ja→Ru or Ru→Ja	12k	-	-	12k	16k
	Ja→En or En→Ja	-	47k	-	47k	16k
	Ru→En or En→Ru	-	-	82k	82k	16k
(a2), (b2)	Ja→Ru and Ru→Ja	12k	-	-	24k	16k
	Ja→En and En→Ja	-	47k	-	94k	16k
	Ru→En and En→Ru	-	-	82k	164k	16k
(a3), (b3)	M2M systems	12k→82k	47k→82k	82k	492k	32k

Table 4: Configuration of uni-, bi-directional, and M2M NMT baseline systems. Arrows in "Parallel data" columns indicate the over-sampling of the parallel data to match the size of the largest parallel data.

unsupervised PBSMT (Lample et al., 2018),[15] it crashed due to unknown reasons. We therefore followed another method described in Marie and Fujita (2018). Instead of short n-grams (Artetxe et al., 2018; Lample et al., 2018), we collected a set of phrases in Japanese and Russian from respective monolingual data using the `word2phrase` algorithm (Mikolov et al., 2013),[16] as in Marie and Fujita (2018). To reduce the complexity, we used randomly selected 10M monolingual sentences, and 300k most frequent phrases made of words among the 300k most frequent words. For each source phrase s, we selected 300-best target phrases t according to the translation probability as in Lample et al. (2018): $p(t|s) = \frac{\exp(\beta \cos(emb(t), emb(s)))}{\sum_{t'} \exp(\beta \cos(emb(t'), emb(s)))}$, where $emb(\cdot)$ stands for a bilingual embedding of a given phrase, obtained through averaging bilingual embeddings of constituent words learned from the two monolingual data using `fastText`[17] and `vecmap`.[18] For each of the retained phrase pair, $p(s|t)$ was computed analogously. We also computed lexical translation probabilities relying on those learned from the given small parallel corpus.

Up to four phrase tables were jointly exploited by the multiple decoding path ability of `Moses`. Weights for the features were tuned using `KB-MIRA` (Cherry and Foster, 2012) on the development set; we took the best weights after 15 iterations. Two hyper-parameters, namely, k for the number of pivot-based phrase pairs per source phrase and d for distortion limit, were determined by a grid search on $k \in \{10, 20, 40, 60, 80, 100\}$ and $d \in \{8, 10, 12, 14, 16, 18, 20\}$. In contrast, we used predetermined hyper-parameters for phrase table induction from monolingual data, following the convention: 200 for the dimension of word and phrase embeddings and $\beta = 30$.

NMT Systems

We used the open-source implementation of the RNMT and the Transformer models in `tensor2tensor`.[19] A uni-directional model for each of the six translation directions was trained on the corresponding parallel corpus. Bi-directional and M2M models were realized by adding an artificial token that specifies the target language to the beginning of each source sentence and shuffling the entire training data (Johnson et al., 2017).

Table 4 contains some specific hyper-parameters[20] for our baseline NMT models. The hyper-parameters not mentioned in this table used the default values in `tensor2tensor`. For M2M systems, we over-sampled Ja→Ru and Ja→En training data so that their sizes match the largest Ru→En data. To reduce the number of unknown words, we used `tensor2tensor`'s internal sub-word segmentation mechanism. Since we work in a low-resource setting, we used shared sub-word vocabularies of size 16k for the uni- and bi-directional models and 32k for the M2M models. The number of training iterations was determined by early-stopping: we evaluated our models on the development set every 1,000 updates, and stopped training if BLEU score for the development set was not improved for 10,000 updates (10 check-points). Note that the development set was created by concatenating those for the individual translation directions without any over-sampling.

Having trained the models, we averaged the last 10 check-points and decoded the test sets with a beam size of 4 and a length penalty which was

[15] `https://github.com/facebookresearch/UnsupervisedMT`
[16] `https://code.google.com/archive/p/word2vec/`
[17] `https://fasttext.cc/`
[18] `https://github.com/artetxem/vecmap`
[19] `https://github.com/tensorflow/tensor2tensor`, version 1.6.6.
[20] We compared two mini-batch sizes, 1024 and 6144 tokens, and found that 6144 and 1024 worked better for RNMT and Transformer, respectively.

ID	System	Ja→Ru	Ru→Ja	Ja→En	En→Ja	Ru→En	En→Ru
(a1)	Uni-directional RNMT	0.58	1.86	2.41	7.83	18.42	13.64
(a2)	Bi-directional RNMT	0.65	1.61	6.18	8.81	19.60	15.11
(a3)	M2M RNMT	1.51	4.29	5.15	7.55	14.24	10.86
(b1)	Uni-directional Transformer	0.70	1.96	4.36	7.97	20.70	16.24
(b2)	Bi-directional Transformer	0.19	0.87	6.48	10.63	22.25	16.03
(b3)	M2M Transformer	**3.72**	**8.35**	**10.24**	**12.43**	22.10	**16.92**
(c1)	Uni-directional supervised PBSMT	2.02	4.45	8.19	10.27	**22.37**	16.52

Table 5: BLEU scores of baseline systems. **Bold** indicates the best BLEU score for each translation direction.

tuned by a linear search on the BLEU score for the development set.

Similarly to PBSMT, we also evaluated "Cascade" and "Synthesize" methods with uni-directional NMT models.

4.3 Results

We evaluated MT models using case-sensitive and tokenized BLEU (Papineni et al., 2002) on test sets, using `Moses`'s *multi-bleu.perl*. Statistical significance ($p < 0.05$) on the difference of BLEU scores was tested by `Moses`'s *bootstrap-hypothesis-difference-significance.pl*.

Tables 5 and 6 show BLEU scores of all the models, except the NMT systems augmented with back-translations. Whereas some models achieved reasonable BLEU scores for Ja↔En and Ru↔En translation, all the results for Ja↔Ru, which is our main concern, were abysmal.

Among the NMT models, Transformer models (b∗) were proven to be better than RNMT models (a∗). RNMT models could not even outperform the uni-directional PBSMT models (c1). M2M models (a3) and (b3) outperformed their corresponding uni- and bi-directional models in most cases. It is worth noting that in this extremely low-resource scenario, BLEU scores of the M2M RNMT model for the largest language pair, i.e., Ru↔En, were lower than those of the uni- and bi-directional RNMT models as in Johnson et al. (2017). In contrast, with the M2M Transformer model, Ru↔En also benefited from multilingualism.

Standard PBSMT models (c1) achieved higher BLEU scores than uni-directional NMT models (a1) and (b1), as reported by Koehn and Knowles (2017), whereas they underperform the M2M Transformer NMT model (b3). As shown in Table 6, pivot-based PBSMT systems always achieved higher BLEU scores than (c1). The best model with three phrase tables, labeled "Synthesize / Triangulate / Gold," brought visible BLEU gains with substantial reduction of OOV tokens (3047→1180 for Ja→Ru, 4463→1812 for Ru→Ja). However, further extension with phrase tables induced from monolingual data did not push the limit, despite their high coverage; only 336 and 677 OOV tokens were left for the two translation directions, respectively. This is due to the poor quality of the bilingual word embeddings used to extract the phrase table, as envisaged in Section 3.

System	Ja→Ru	Ru→Ja
PBSMT: Cascade	3.65	7.62
PBSMT: Synthesize	3.37	6.72
PBSMT: Synthesize / Gold	2.94	6.95
PBSMT: Synthesize + Gold	3.07	6.62
PBSMT: Triangulate	**3.75**	7.02
PBSMT: Triangulate / Gold	**3.93**	7.02
PBSMT: Synthesize / Triangulate / Gold	**4.02**	7.07
PBSMT: Induced	0.37	0.65
PBSMT: Induced / Synthesize / Triangulate / Gold	2.85	6.86
RNMT: Cascade	1.19	6.73
RNMT: Synthesize	1.82	3.02
RNMT: Synthesize + Gold	1.62	3.24
Transformer NMT: Cascade	2.41	6.84
Transformer NMT: Synthesize	1.78	5.43
Transformer NMT: Synthesize + Gold	2.13	5.06

Table 6: BLEU scores of pivot-based systems. "Gold" refers to the phrase table trained on the parallel data. **Bold** indicates the BLEU score higher than the best one in Table 5. "/" indicates the use of separately trained multiple phrase tables, whereas so does "+" training on the mixture of parallel data.

None of pivot-based approaches with uni-directional NMT models could even remotely rival the M2M Transformer NMT model (b3).

4.4 Augmentation with Back-translation

Given that the M2M Transformer NMT model (b3) achieved best results for most of the translation directions and competitive results for the rest, we further explored it through back-translation.

We examined the utility of pseudo-parallel data for all the six translation directions, unlike the work of Lakew et al. (2017) and Lakew et al. (2018), which concentrate only on the zero-shot language pair, and the work of Niu et al. (2018), which compares only uni- or bi-directional models. We investigated whether each translation direction in M2M models will benefit from pseudo-parallel data and if so, what kind of improvement takes place.

ID	System	Parallel data				Total size of training data
		Pseudo	Ja↔Ru	Ja↔En	Ru↔En	
#1–#10	Ja∗→Ru and/or Ru∗→Ja	12k→82k	12k→82k	47k→82k×2	82k×2	984k
	Ja∗→En and/or En∗→Ja	47k→82k	12k→82k×2	47k→82k	82k×2	984k
	Ru∗→En and/or En∗→Ru	82k	12k→82k×2	47k→82k×2	82k	984k
	All	All of the above	12k→82k	47k→82k	82k	984k

Table 7: Over-sampling criteria for pseudo-parallel data generated by back-translation.

ID	Pseudo-parallel data involved						BLEU score					
	Ja∗→Ru	Ru∗→Ja	Ja∗→En	En∗→Ja	Ru∗→En	En∗→Ru	Ja→Ru	Ru→Ja	Ja→En	En→Ja	Ru→En	En→Ru
(b3)	-	-	-	-	-	-	3.72	8.35	10.24	12.43	22.10	16.92
#1	✓	-	-	-	-	-	**•4.59**	**8.63**	**10.64**	**12.94**	**22.21**	**17.30**
#2	-	✓	-	-	-	-	**3.74**	**•8.85**	10.13	**13.05**	**22.48**	**17.20**
#3	✓	✓	-	-	-	-	**•4.56**	**•9.09**	**10.57**	**•13.23**	**22.48**	**•17.89**
#4	-	-	✓	-	-	-	3.71	8.05	**•11.00**	**12.66**	**22.17**	16.76
#5	-	-	-	✓	-	-	3.62	8.10	9.92	**•14.06**	21.66	16.68
#6	-	-	✓	✓	-	-	3.61	7.94	**•11.51**	**•14.38**	**22.22**	16.80
#7	-	-	-	-	✓	-	**3.80**	**8.37**	**10.67**	**13.00**	**22.51**	**•17.73**
#8	-	-	-	-	-	✓	**3.77**	8.04	**10.52**	12.43	**•22.85**	**17.13**
#9	-	-	-	-	✓	✓	3.37	8.03	10.19	**12.79**	**22.77**	**17.26**
#10	✓	✓	✓	✓	✓	✓	**•4.43**	**•9.38**	**•12.06**	**•14.43**	**•23.09**	**17.30**

Table 8: BLEU scores of M2M Transformer NMT systems trained on the mixture of given parallel corpus and pseudo-parallel data generated by back-translation using (b3). Six "X∗→Y" columns show whether the pseudo-parallel data for each translation direction is involved. **Bold** indicates the scores higher than (b3) and "•" indicates statistical significance of the improvement.

First, we selected sentences to be back-translated from in-domain monolingual data (Table 3), relying on the score proposed by Moore and Lewis (2010) via the following procedure.

1. For each language, train two 4-gram language models, using KenLM: an in-domain one on all the Global Voices data, i.e., both parallel and monolingual data, and a general-domain one on the concatenation of Global Voices, IWSLT, and Tatoeba data.
2. For each language, discard sentences containing OOVs according to the in-domain language model.
3. For each translation direction, select the T-best monolingual sentences in the news domain, according to the difference between cross-entropy scores given by the in-domain and general-domain language models.

Whereas Niu et al. (2018) exploited monolingual data much larger than parallel data, we maintained a 1:1 ratio between them (Johnson et al., 2017), setting T to the number of lines of parallel data of given language pair.

Selected monolingual sentences were then translated using the M2M Transformer NMT model (b3) to compose pseudo-parallel data. Then, the pseudo-parallel data were enlarged by over-sampling as summarized in Table 7. Finally, new NMT models were trained on the concatenation of the original parallel and pseudo-parallel data from scratch in the same manner as the previous NMT models with the same hyper-parameters.

Table 8 shows the BLEU scores achieved by several reasonable combinations of six-way pseudo-parallel data. We observed that the use of all six-way pseudo-parallel data (#10) significantly improved the base model for all the translation directions, except En→Ru. A translation direction often benefited when the pseudo-parallel data for that specific direction was used.

4.5 Summary

We have evaluated an extensive variation of MT models[21] that rely only on in-domain parallel and monolingual data. However, the resulting BLEU scores for Ja→Ru and Ru→Ja tasks do not exceed 10 BLEU points, implying the inherent limitation of the in-domain data as well as the difficulty of these translation directions.

5 Exploiting Large Out-of-Domain Data Involving a Helping Language

The limitation of relying only on in-domain data demonstrated in Section 4 motivates us to explore

[21] Other conceivable options include transfer learning using parallel data between English and one of Japanese and Russian as either source or target language, such as pre-training an En→Ru model and fine-tuning it for Ja→Ru. Our M2M models conceptually subsume them, even though they do not explicitly divide the two steps during training. On the other hand, our method proposed in Section 5 explicitly conducts transfer learning for domain adaptation followed by additional transfer learning across different languages.

Domain \ language pair	Direct	One-side shared
in-domain	A, ✓	B, ✓
out-of-domain	C, ×	D, ✓

Table 9: Classification of parallel data.

Lang.pair	Corpus	#sent.	#tokens	#types
Ja↔En	ASPEC	1,500,000	42.3M / 34.6M	234k / 1.02M
Ru↔En	UN	2,647,243	90.5M / 92.8M	757k / 593k
	Yandex	320,325	8.51M / 9.26M	617k / 407k

Table 10: Statistics on our out-of-domain parallel data.

other types of parallel data. As raised in our second research question, [RQ2], we considered the effective ways to exploit out-of-domain data.

According to language pair and domain, parallel data can be classified into four categories in Table 9. Among all the categories, out-of-domain data for the language pair of interest have been exploited in the domain adaptation scenarios (C→A) (Chu et al., 2017). However, for Ja↔Ru, no out-of-domain data is available. To exploit out-of-domain parallel data for Ja↔En and Ru↔En pairs instead, we propose a multistage fine-tuning method, which combines two types of transfer learning, i.e., domain adaptation for Ja↔En and Ru↔En (D→B) and multilingual transfer (B→A), relying on the M2M model examined in Section 4. We also examined the utility of fine-tuning for iteratively generating and using pseudo-parallel data.

5.1 Multistage Fine-tuning

Simply using NMT systems trained on out-of-domain data for in-domain translation is known to perform badly. In order to effectively use large-scale out-of-domain data for our extremely low-resource task, we propose to perform domain adaptation through either (a) conventional fine-tuning, where an NMT system trained on out-of-domain data is fine-tuned only on in-domain data, or (b) mixed fine-tuning (Chu et al., 2017), where pre-trained out-of-domain NMT system is fine-tuned using a mixture of in-domain and out-of-domain data. The same options are available for transferring from Ja↔En and Ru↔En to Ja↔Ru.

We inevitably involve two types of transfer learning, i.e., domain adaptation for Ja↔En and Ru↔En and multilingual transfer for Ja↔Ru pair. Among several conceivable options for managing these two problems, we examined the following multistage fine-tuning.

Stage 0. Out-of-domain pre-training: Pre-train a multilingual model only on the Ja↔En and Ru↔En out-of-domain parallel data (I), where the vocabulary of the model is determined on the basis of the in-domain parallel data in the same manner as the M2M NMT models examined in Section 4.

Stage 1. Fine-tuning for domain adaptation: Fine-tune the pre-trained model (I) on the in-domain Ja↔En and Ru↔En parallel data (fine-tuning, II) or on the mixture of in-domain and out-of-domain Ja↔En and Ru↔En parallel data (mixed fine-tuning, III).

Stage 2. Fine-tuning for Ja↔Ru pair: Further fine-tune the models (each of II and III) for Ja↔Ru on in-domain parallel data for this language pair only (fine-tuning, IV and VI) or on all the in-domain parallel data (mixed fine-tuning, V and VII).

We chose this way due to the following two reasons. First, we need to take a balance between several different parallel corpora sizes. The other reason is division of labor; we assume that solving each sub-problem one by one should enable gradual shift of parameters.

5.2 Data Selection

As an additional large-scale out-of-domain parallel data for Ja↔En, we used the cleanest 1.5M sentences from the Asian Scientific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2016).[22] As for Ru↔En, we used the UN and Yandex corpora released for the WMT 2018 News Translation Task.[23] We retained Ru↔En sentence pairs that contain at least one OOV token in both sides, according to the in-domain language model trained in Section 4.4. Table 10 summarizes the statistics on the remaining out-of-domain parallel data.

5.3 Results

Table 11 shows the results of our multistage fine-tuning, where the IDs of each row refer to those described in Section 5.1. First of all, the final models of our multistage fine-tuning, i.e., V and VII, achieved significantly higher BLEU scores than (b3) in Table 5, a weak baseline without using any monolingual data, and #10 in Table 8, a strong baseline established with monolingual data.

[22] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[23] http://www.statmt.org/wmt18/translation-task.html

ID	Initialized	Out-of-domain data		In-domain data			BLEU score					
		Ja↔En	Ru↔En	Ja↔Ru	Ja↔En	Ru↔En	Ja→Ru	Ru→Ja	Ja→En	En→Ja	Ru→En	En→Ru
(b3)	-	-	-	✓	✓	✓	3.72	8.35	10.24	12.43	22.10	16.92
I	-	✓	✓	-	-	-	0.00	0.15	4.59	4.15	•25.22	•20.37
II	I	-	-	-	✓	✓	0.20	0.70	•14.10	•**17.80**	•28.23	•24.35
III	I	✓	✓	-	✓	✓	0.23	1.07	•13.31	•17.74	•**28.73**	•**25.22**
IV	II	-	-	✓	-	-	•5.44	•10.67	0.12	3.97	0.11	3.66
V	II	-	-	✓	✓	✓	•6.90	•11.99	•14.34	•16.93	•27.50	•23.17
VI	III	-	-	✓	-	-	•5.91	•10.83	0.26	2.18	0.18	1.10
VII	III	-	-	✓	✓	✓	•7.49	•12.10	•**14.63**	•17.51	•28.51	•24.60
I'	-	✓	✓	✓	✓	✓	•5.31	•10.73	•14.41	•16.34	•27.46	•23.21
II'	I	-	-	✓	✓	✓	•6.30	•11.64	•14.29	•16.83	•27.53	•23.00
III'	I	✓	✓	✓	✓	✓	•**7.53**	•**12.33**	•14.19	•16.77	•27.94	•23.97

Table 11: BLEU scores obtained through multistage fine-tuning. "Initialized" column indicates the model used for initializing parameters that are fine-tuned on the data indicated by ✓. **Bold** indicates the best BLEU score for each translation direction. "•" indicates statistical significance of the improvement over (b3).

The performance of the initial model (I) depends on the language pair. For Ja↔Ru pair, it cannot achieve minimum level of quality since the model has never seen parallel data for this pair. The performance on Ja↔En pair was much lower than the two baseline models, reflecting the crucial mismatch between training and testing domains. In contrast, Ru↔En pair benefited the most and achieved surprisingly high BLEU scores. The reason might be due to the proximity of out-of-domain training data and in-domain test data.

The first fine-tuning stage significantly pushed up the translation quality for Ja↔En and Ru↔En pairs, in both cases with fine-tuning (II) and mixed fine-tuning (III). At this stage, both models performed only poorly for Ja↔Ru pair as they have not yet seen Ja↔Ru parallel data. Either model had a consistent advantage to the other.

When these models were further fine-tuned only on the in-domain Ja↔Ru parallel data (IV and VI), we obtained translations of better quality than the two baselines for Ja↔Ru pair. However, as a result of complete ignorance of Ja↔En and Ru↔En pairs, the models only produced translations of poor quality for these language pairs. In contrast, mixed fine-tuning for the second fine-tuning stage (V and VII) resulted in consistently better models than conventional fine-tuning (IV and VI), irrespective of the choice at the first stage, thanks to the gradual shift of parameters realized by in-domain Ja↔En and Ru↔En parallel data. Unfortunately, the translation quality for Ja↔En and Ru↔En pairs sometimes degraded from II and III. Nevertheless, the BLEU scores still retain the large margin against two baselines.

The last three rows in Table 11 present BLEU scores obtained by the methods with fewer fine-tuning steps. The most naive model I', trained on the balanced mixture of whole five types of corpora from scratch, and the model II', obtained through a single-step conventional fine-tuning of I on all the in-domain data, achieved only BLEU scores consistently worse than VII. In contrast, when we merged our two fine-tuning steps into a single mixed fine-tuning on I, we obtained a model III' which is better for the Ja↔Ru pair than VII. Nevertheless, they are still comparable to those of VII and the BLEU scores for the other two language pairs are much lower than VII. As such, we conclude that our multistage fine-tuning leads to a more robust in-domain multilingual model.

5.4 Further Augmentation with Back-translation

Having obtained a better model, we examined again the utility of back-translation. More precisely, we investigated (a) whether the pseudo-parallel data generated by an improved NMT model leads to a further improvement, and (b) whether one more stage of fine-tuning on the mixture of original parallel and pseudo-parallel data will result in a model better than training a new model from scratch as examined in Section 4.4.

Given an NMT model, we first generated six-way pseudo-parallel data by translating monolingual data. For the sake of comparability, we used the identical monolingual sentences sampled in Section 4.4. Then, we further fine-tuned the given model on the mixture of the generated pseudo-parallel data and the original parallel data, following the same over-sampling procedure in Section 4.4. We repeated these steps five times.

Table 12 shows the results. "new #10" in the second row indicates an M2M Transformer model trained from scratch on the mixture of six-way pseudo-parallel data generated by VII and the orig-

No	Initialized	BT	BLEU score					
			Ja→Ru	Ru→Ja	Ja→En	En→Ja	Ru→En	En→Ru
#10	-	(b3)	4.43	9.38	12.06	14.43	23.09	17.30
new #10	-	VII	•6.55	•11.36	•13.77	•15.59	•24.91	•20.55
VIII	VII	VII	•7.83	•12.21	•15.06	•17.19	•28.49	•23.96
IX	VIII	VIII	•8.03	•12.55	•15.07	•17.80	•28.16	•24.27
X	IX	IX	•7.76	•12.59	•15.08	•18.12	•28.18	•24.67
XI	X	X	•7.85	•12.97	•15.26	•17.83	•28.49	•24.36
XII	XI	XI	•8.16	•13.09	•14.96	•17.74	•28.45	•24.35

Table 12: BLEU scores achieved through fine-tuning on the mixture of the original parallel data and six-way pseudo-parallel data. "Initialized" column indicates the model used for initializing parameters and so does "BT" column the model used to generate pseudo-parallel data. "•" indicates statistical significance of the improvement over #10.

Investigation step	Ja→Ru	Ru→Ja
Uni-directional Transformer: (b1) in Table 5	0.70	1.96
M2M Transformer: (b3) in Table 5	3.72	8.35
+ six-way pseudo-parallel data: #10 in Table 8	4.43	9.38
M2M multistage fine-tuning: VII in Table 11	7.49	12.10
+ six-way pseudo-parallel data: XII in Table 12	8.16	13.09

Table 13: Summary of our investigation: BLEU scores of the best NMT systems at each step.

inal parallel data. It achieved higher BLEU scores than #10 in Table 8 thanks to the pseudo-parallel data of better quality, but underperformed the base NMT model VII. In contrast, our fine-tuned model VIII successfully surpassed VII, and one more iteration (IX) further improved BLEU scores for all translation directions, except Ru→En. Although further iterations did not necessarily gain BLEU scores, we came to a much higher plateau compared to the results in Section 4.

6 Conclusion

In this paper, we challenged the difficult task of Ja↔Ru news domain translation in an extremely low-resource setting. We empirically confirmed the limited success of well-established solutions when restricted to in-domain data. Then, to incorporate out-of-domain data, we proposed a multilingual multistage fine-tuning approach and observed that it substantially improves Ja↔Ru translation by over 3.7 BLEU points compared to a strong baseline, as summarized in Table 13. This paper contains an empirical comparison of several existing approaches and hence we hope that our paper can act as a guideline to researchers attempting to tackle extremely low-resource translation.

In the future, we plan to confirm further fine-tuning for each of specific translation directions. We will also explore the way to exploit out-of-domain pseudo-parallel data, better domain-adaptation approaches, and additional challenging language pairs.

Acknowledgments

This work was carried out when Aizhan Imankulova was taking up an internship at NICT, Japan. We would like to thank the reviewers for their insightful comments. A part of this work was conducted under the program "Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology" of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA.

Cherry, Colin and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada.

Cho, Kyunghyun, Bart van Merriënboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar.

Chu, Chenhui, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 385–391, Vancouver, Canada.

Cohn, Trevor and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic.

Firat, Orhan, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, USA.

Johnson, Howard, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 967–975, Prague, Czech Republic.

Johnson, Melvin, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kocmi, Tom and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Lakew, Surafel M, Quintino F. Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Improving zero-shot translation of low-resource languages. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 113–119, Tokyo, Japan.

Lakew, Surafel M, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, USA.

Lample, Guillaume, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium.

Marie, Benjamin and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *CoRR*, abs/1810.12703.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 3111–3119, Lake Tahoe, USA. Curran Associates Inc.

Moore, Robert C. and Will Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL) Short Papers*, pages 220–224, Uppsala, Sweden.

Nakazawa, Toshiaki, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2204–2208, Portorož, Slovenia.

Niu, Xing, Michael Denkowski, and Marine Carpuat. 2018. Bi-directional neural machine translation with synthetic parallel data. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 84–91, Melbourne, Australia.

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

Sennrich, Rico, 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, Berlin, Germany.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 3104–3112, Montréal, Canada.

Tillmann, Christoph and Jian-ming Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 93–96, Boulder, USA.

Utiyama, Masao and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, Sapporo, Japan.

Utiyama, Masao and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, USA.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of 30th Advances in Neural Information Processing Systems*, pages 5998–6008.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, USA.