Zero-Resource Neural Machine Translation with Monolingual Pivot Data

Anna Currey
University of Edinburgh
a.currey@sms.ed.ac.uk

Kenneth Heafield
University of Edinburgh
kheafiel@ed.ac.uk

Abstract

Zero-shot neural machine translation (NMT) is a framework that uses source-pivot and target-pivot parallel data to train a source-target NMT system. An extension to zero-shot NMT is zero-resource NMT, which generates pseudo-parallel corpora using a zero-shot system and further trains the zero-shot system on that data. In this paper, we expand on zero-resource NMT by incorporating monolingual data in the pivot language into training; since the pivot language is usually the highest-resource language of the three, we expect monolingual pivot-language data to be most abundant. We propose methods for generating pseudo-parallel corpora using pivot-language monolingual data and for leveraging the pseudo-parallel corpora to improve the zero-shot NMT system. We evaluate these methods for a high-resource language pair (German-Russian) using English as the pivot. We show that our proposed methods yield consistent improvements over strong zero-shot and zero-resource baselines and even catch up to pivot-based models in BLEU (while not requiring the two-pass inference that pivot models require).

1 Introduction

Neural machine translation (NMT) has achieved impressive results on several high-resource translation tasks (Hassan et al., 2018; Wu et al., 2016). However, these systems have relied on large amounts of parallel training data between the source and the target language; for many language pairs, such data may not be available. Even two high-resource languages, such as German and Russian, may not have sufficient parallel data between them.

Recently, unsupervised NMT systems that learn to translate using only monolingual corpora have been proposed as a solution to this problem (Artetxe et al., 2018; Lample et al., 2018). However, such systems do not make full use of available parallel corpora between the source and target languages and a potential pivot language.

Although most language pairs may have little in-domain parallel data available, it is often possible to find parallel corpora with a third *pivot* language. For example, while German↔Russian parallel data is relatively scarce, German↔English and Russian↔English data is abundant. Pivot-based and zero-shot NMT systems have been proposed as a means of taking advantage of this data to translate between e.g. German and Russian.

In pivot-based machine translation, text is first translated from the source language into the pivot language, and then from the pivot language into the target language. Although such methods can result in strong translation performance (Johnson et al., 2017), they have a few disadvantages. The two-step pivoting translation process doubles the latency during inference and has the potential to propagate errors from the source→pivot translation into the final target output. Additionally, there is a risk that relevant information in the source sentence can be lost in the pivot translation (e.g. case distinctions if pivoting through English) and not represented in the target sentence. Zero-shot methods that take advantage of multilingual NMT systems to perform direct source→target translation have become a popular method for addressing this problem, and zero-resource methods build off of zero-shot methods by fine-tuning on pseudo-parallel data to improve direct translation (see section 2.1 for a review of zero-shot and zero-resource methods). Zero-resource methods are beneficial because they can potentially take advantage of all available training data, including parallel and monolingual corpora.

The goal of this paper is to augment zero-resource NMT with monolingual data from the

Proceedings of the 3rd Workshop on Neural Generation and Translation (WNGT 2019), pages 99–107

www.aclweb.org/anthology/D19-56%2d

pivot language. Although there have been several explorations into using parallel corpora through a pivot language to improve NMT (Firat et al., 2016; Lakew et al., 2017; Park et al., 2017) and using monolingual source and target corpora in NMT (Edunov et al., 2018; Gulcehre et al., 2015; Hoang et al., 2018; Niu et al., 2018; Sennrich et al., 2016a; Zhang and Zong, 2016), this is to our knowledge the first attempt at using monolingual pivot-language data to augment NMT training. Leveraging monolingual pivot-language data is worthwhile because the pivot language is often the highest-resource language of the three (e.g. it is often English), so we expect there to be more high-quality monolingual pivot data than monolingual source or target data in many cases. Thus, we make use of parallel source↔pivot data, parallel target↔pivot data, and monolingual pivot-language data to build a zero-resource NMT system. Although we use a basic multilingual NMT system as the basis, the methods proposed here could easily be applied to any zero-shot NMT architecture.

2 Related Work

2.1 Zero-Shot and Zero-Resource NMT

Zero-shot neural machine translation, i.e. NMT between two languages for which no parallel data was used at training time, is often done by leveraging multilingual NMT systems. Firat et al. (2016) first attempted zero-shot NMT with a multilingual model consisting of several encoders and decoders, but found that without fine-tuning, the model was not able to translate between the zero-shot language pairs. On the other hand, multilingual NMT with shared encoders and decoders (Ha et al., 2016; Johnson et al., 2017) is more successful at zero-shot NMT, although its performance still lags behind pivoting.

Several modifications to the multilingual NMT architecture have been proposed with the goal of improving zero-shot NMT performance; here, we review some such modifications. Lu et al. (2018) added an interlingua layer to the multilingual NMT model; this layer transforms language-specific encoder outputs into language-independent decoder inputs. Platanios et al. (2018) updated the shared encoder/decoder multilingual NMT model by adding a contextual parameter generator. This generator generates the encoder and decoder parameters for a given source and target language, taking only source and target language as input. Arivazhagan et al. (2019) augmented the NMT loss function with a term that promotes the creation of an interlingua.

In this paper, we concentrate on the task of zero-resource translation, which starts from a multilingual NMT system and improves the zero-shot direction using pseudo-parallel corpora. Firat et al. (2016) found that zero-shot NMT performance could be strongly improved by fine-tuning on a pseudo-parallel corpus created by back-translating from the pivot language into each zero-shot language. Similarly, Lakew et al. (2017) improved low-resource zero-shot NMT by back-translating directly between the two zero-shot languages and fine-tuning on the resulting corpus. Park et al. (2017) combined both of these methods and also included NMT-generated sentences on the target side of the pseudo-parallel corpora.

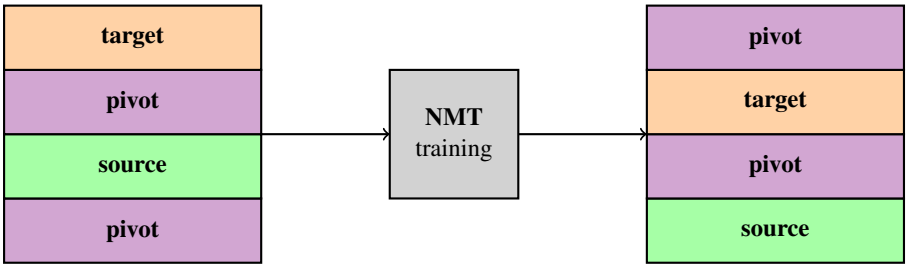
2.2 NMT with Monolingual Data

This paper builds off the idea of back-translation in order to incorporate pivot-language monolingual data into NMT. Back-translation was introduced for NMT by Sennrich et al. (2016a). This technique consists of first training a target→source NMT system and using that to translate the target monolingual data into the source language. The resulting pseudo-parallel source→target corpus is used to augment the training of the final system.

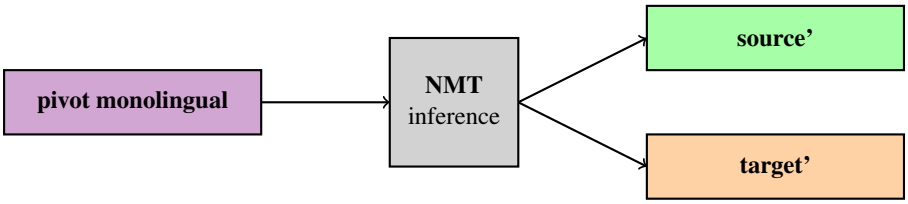
Several methods for improving back-translation have also been proposed. Zhang and Zong (2016) extended back-translation to monolingual source-language data by using the initial system to translate the source data to the target and re-training on the resulting pseudo-parallel corpus. Niu et al. (2018) augmented multilingual NMT with back-translation. They trained a single model for source→target and target→source translation, used that model to back-translate source and target monolingual data, and fine-tuned the model on the back-translated corpora.

3 Zero-Resource NMT with Pivot Monolingual Data

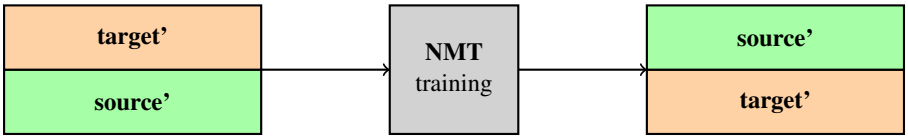
In this paper, we concentrate on zero-resource NMT between two languages X and Y given a pivot language Z. We assume access to X↔Z and Y↔Z parallel corpora, but no direct X↔Y parallel corpus. Our goal is to use additional monolingual data in the pivot language Z to improve both



(a) An initial multilingual NMT model is trained on source↔pivot and target↔pivot parallel data (section 3.1).



(b) The pivot monolingual corpus is back-translated into the source and target languages using the trained NMT model (section 3.2).



(c) The source'→target' and target'→source' pseudo-parallel corpora are used to train the final NMT system from scratch or fine-tune the initial model (section 3.3). In practice, we concatenate this data with a subset of the original parallel data (not shown here).

Figure 1: Illustration of the basic steps in our zero-resource NMT model using pivot-language monolingual data.

X→Y and Y→X translation simultaneously. Figure 1 gives an overview of our proposed method.

3.1 Initial Multilingual Models

We start by giving an overview of the multilingual NMT models that are used as the basis for our experiments. Here, we do not consider single-directional bilingual NMT models, only multilingual NMT models. This is because we would like to translate directly between language X and language Y at inference time without using the pivot language; translating through the pivot language would double the amount of time it takes to translate and potentially lead to information loss or error propagation. In this work, we also do not consider the case of adding monolingual data from the main languages of interest (X and Y), although such data would likely further improve translation quality.

Our initial multilingual NMT model is based on the model introduced by Johnson et al. (2017), although here we use the transformer architecture (Vaswani et al., 2017). We train the initial model on mixed X→Z, Z→X, Y→Z, and Z→Y parallel data and use tags at the beginning and end of each source sentence to indicate the desired target language. We shuffle all of the data together randomly, regardless of source and target language. We do not employ any extensions to the zero-shot architecture (Arivazhagan et al., 2019; Lu et al., 2018; Platanios et al., 2018), although the methods described here could easily be applied to such extensions as well.

3.2 Back-Translation of Pivot Monolingual Data

We turn now to the task of leveraging the monolingual corpus in the pivot language Z to improve the multilingual NMT models. We aim to improve only X→Y and Y→X translation, without regard to performance on the other language pairs that are included in the multilingual system (X↔Z and Y↔Z).

First, we use the initial multilingual model described in section 3.1 to back-translate the monolingual pivot data into both languages of interest (X and Y). Since the initial multilingual model was trained on both these directions (Z→X and

Method	Back-Translated Data	Training Regime
pivot from scratch	BT-pivot	train from scratch
pivot fine-tune	BT-pivot	fine-tune initial model
pivot-parallel combined	BT-pivot + BT-parallel	fine-tune initial model

Table 1: Summary of the proposed methods for zero-shot NMT using pivot-language monolingual data.

Z→Y), we expect it to do reasonably well at back-translation. Thus, for each sentence in the Z monolingual corpus, we have its translation in both X and Y, so we can create a pseudo-parallel corpus X'↔Y' (where the prime symbol indicates machine-translated text). We concatenate both directions (X'→Y' and Y'→X') together to create our back-translated pivot (*BT-pivot*) corpus. This resulting corpus contains synthetic data on both the source and the target side.

3.3 Using the BT-Pivot Corpus

The BT-pivot corpus uses the monolingual corpus from the pivot language Z to create a direct pseudo-parallel corpus between the two languages of interest, X and Y. In this section, we introduce three methods for using this BT-pivot data to create a zero-resource NMT system for X↔Y translation. In all cases, we concatenate the BT-pivot corpus with a subset of the original training data to train the zero-resource models; in preliminary experiments, we found that using some original training data yielded slightly higher BLEU scores than training on back-translated data alone. We take only a subset of the original parallel training data rather than the entire corpus in order to cut down on training time.

We dub our first method *pivot from scratch*. In this method, we discard the initial NMT model and train a new NMT model from scratch using the BT-pivot data (concatenated with the subset of the original parallel corpora). We use the same model hyperparameters as for the initial NMT model.

Our second method, *pivot fine-tune*, is similar to the first: both methods use the BT-pivot data (along with the subset of the original parallel data). However, for pivot fine-tune, we use the BT-pivot data and the subset of the parallel data to fine-tune the original multilingual model described in section 3.1, rather than training a new model from scratch.

Finally, we propose a *pivot-parallel combined* method. This method also fine-tunes the original multilingual model, but uses an augmented fine-tuning dataset. In addition to the BT-pivot corpus and the subset of the original training data, we add a back-translated parallel (*BT-parallel*) corpus generated following Firat et al. (2016) as follows:

1. Use the initial multilingual model to translate the Z side of the subsetted X↔Z parallel corpus into language Y.
2. Combine the resulting Y' data with the X side of the subsetted X↔Z parallel corpus to create a Y'→X parallel corpus.
3. Use the initial multilingual model to translate the Z side of the subsetted Y↔Z parallel corpus into language X.
4. Combine the resulting X' data with the Y side of the subsetted Y↔Z parallel corpus to create a X'→Y parallel corpus.
5. Concatenate the two back-translated corpora (X'→Y and Y'→X) to create the BT-parallel corpus.

The BT-parallel corpus is then combined with the BT-pivot corpus and the subset of the original parallel data and used to fine-tune the initial multilingual model.

Table 1 summarizes the three proposed methods for zero-shot NMT. The three methods vary in the back-translated data used (BT-pivot only vs. BT-pivot and BT-parallel) and in the training regime (training a new model from scratch vs. fine-tuning the initial multilingual model). In initial experiments, we also tried a version of the pivot-parallel combined method that trained a new model from scratch, although this did not do as well as the pivot-parallel combined method with fine-tuning.

4 Experimental Setup

4.1 Data

We run our experiments on a high-resource setting: translation between German (DE) and Russian (RU) using English (EN) as the pivot. The data comes from the WMT16 news translation

Corpus	Sentences
EN↔DE	4 497 878
EN↔RU	2 500 502
EN monolingual	1 000 000

Table 2: Number of sentences in each training corpus for the DE↔RU experiments.

task (Bojar et al., 2016). We use all available parallel corpora for EN↔DE (Europarl v7, Common Crawl, and News Commentary v11) and for EN↔RU (Common Crawl, News Commentary v11, Yandex Corpus, and Wiki Headlines) to train the initial multilingual system, but no direct DE↔RU parallel data. When the parallel data is used alongside the back-translated corpora for fine-tuning or re-training from scratch (as described in section 3.1), we randomly sample one million sentences from each parallel corpus.

For pivot (EN) monolingual data, we take a random subset of one million sentences from the News Crawl 2015 corpus. Since the goal of this paper is to study the effectiveness of using pivot-language monolingual data, we do not use any DE or RU monolingual data; however, we expect that such data would also be beneficial. Table 2 shows the size of each training corpus after preprocessing. We use the overlapping DE and RU sentences from newstest2014 as the validation set (1505 sentences), newstest2015 as the test set (1433 sentences), and newstest2016 as the held-out set (1500 sentences). The overlapping sentences were originally written in English and were translated by human translators into German and Russian (Bojar et al., 2016).

All data is tokenized and truecased using the Moses scripts (Koehn et al., 2007). We use a joint byte pair encoding (Sennrich et al., 2016b) vocabulary for all three languages (DE, EN, and RU) trained on all parallel data with 50k merge operations. Similarly to Johnson et al. (2017), we use tags at the beginning and end of the source sentence to indicate the desired target language.

4.2 Models

All models in our experiments are based on the transformer architecture (Vaswani et al., 2017). We use the Sockeye toolkit (Hieber et al., 2017) to run all experiments. We find that the default Sockeye hyperparameters work well, so we stick with those throughout. We use beam search with beam size 5 both when back-translating and during inference.

4.3 Baselines

Initial Models Without Monolingual Data

We compare our models to three baselines that are trained without any monolingual data. We refer to these baselines as *initial models* because they are used as the basis for our proposed models: we use them to generate the BT-pivot data and we fine-tune them using the generated data to create our proposed models.

The first baseline is a multilingual model based on Johnson et al. (2017), but we use the transformer architecture and add target language tags at both the beginning and end of the source sentences. This multilingual model is trained on the English↔German and English↔Russian parallel data. We evaluate this model both with direct (zero-shot) translation (German→Russian and Russian→German) and with pivot translation through English.

Secondly, we consider the zero-resource NMT method proposed by Lakew et al. (2017). This method consists of selecting sentences from the DE↔EN parallel corpus and back-translating them from DE into RU, resulting in a RU'→DE pseudo-parallel corpus. The same is also done with the RU↔EN parallel corpus to create a DE'→RU pseudo-parallel corpus. These corpora are then concatenated with the original parallel data and used to fine-tune the multilingual model. This zero-resource method is only evaluated on direct DE→RU and RU→DE translation (not on pivoting through EN).

We also compare our models to a zero-resource baseline based on the technique introduced by Firat et al. (2016). This method fine-tunes the initial multilingual model with the BT-parallel corpus described in section 3.3 (concatenated with the origial data). Like the other zero-resource baseline, this baseline is only evaluated on direct translation (not on pivot translation).

Baselines with Monolingual Data

In addition to the initial models, we compare our proposed zero-resource NMT methods to two baselines trained with monolingual EN data. For both of these baselines, we evaluate both direct zero-shot translation and pivot translation through EN.

	BLEU	RU→DE		DE→RU	
		test	held-out	test	held-out
initial models	multilingual direct	15.2	14.5	3.4	2.7
	multilingual pivot	21.7	20.2	21.3	19.3
	Lakew et al., 2017	14.4	13.2	19.4	17.0
	Firat et al., 2016	21.0	18.3	22.6	20.7
baselines	copied corpus direct	10.2	9.5	3.7	3.1
	copied corpus pivot	21.1	19.9	20.9	18.9
	back-translation direct	14.8	14.1	3.7	2.9
	back-translation pivot	22.4	20.9	22.3	20.4
proposed models	pivot from scratch	22.3	21.5	23.0	20.6
	pivot fine-tune	22.4	21.5	23.0	20.3
	pivot-parallel combined	**22.5**	**21.6**	**23.6**	**21.1**

Table 3: BLEU scores for the initial multilingual models and zero-resource models without monolingual data, for the baselines with pivot monolingual data, and for our proposed zero-resource models with pivot monolingual data. We report results on the test set (newstest2015) and the held-out set (newstest2016). For the baselines and the initial multilingual models, we use consider both direct (zero-shot) and pivot translation.

The first is based on the copied corpus method of Currey et al. (2017). We train an identical model to the initial multilingual model, but with additional EN→EN pseudo-parallel training data from the EN monolingual corpus. Thus, this model is trained on DE↔EN, RU↔EN, and EN→EN data. We do not fine-tune this model with any pseudo-parallel data.

The second baseline we consider is back-translation (Sennrich et al., 2016a). Starting from the trained multilingual model, we back-translate the EN monolingual data into both DE and RU, then fine-tune the multilingual model on the original training data, plus the DE'→EN and RU'→EN pseudo-parallel corpora.

5 Results

Table 3 shows translation performance (as estimated by BLEU score) for our main experiments. We display results for initial multilingual models without any monolingual data (rows 1–4), for copied corpus and back-translation baselines using the monolingual data (rows 5–8), and for our proposed zero-resource models (rows 9–11). For the initial multilingual model and for the copied corpus and back-translation baselines, we consider both direct source→target translation and translation through the pivot language (source→EN→target).

5.1 Initial Models Without Monolingual Data

For the multilingual baseline, direct source→target translation does very poorly for DE→RU. Although the performance is somewhat more reasonable for RU→DE, direct translation still lags far behind pivot (source→EN→target) translation for this model. Our results differ from those of Johnson et al. (2017), who showed reasonable performance in both directions for zero-shot translation. However, they tested their zero-shot systems only on closely related languages or very large-scale multilingual systems, whereas we use somewhat smaller training sets and distantly related languages. This might be an explanation for the discrepancy in results.

Both zero-resource models (Lakew et al., 2017 and Firat et al., 2016) outperform the multilingual baseline overall for direct translation. In addition, the latter closes the gap with the pivot translation baseline for DE→RU and almost closes it for RU→DE. Thus, fine-tuning on back-translated parallel data is very helpful in improving zero-resource NMT. In the next sections, we evaluate methods for further improving zero-resource NMT using EN monolingual data.

5.2 Baselines with Monolingual Data

The results for the copied corpus and back-translation baselines (using both direct and pivot translation) are shown in rows 5–8 of Table 3. Both models are unable to translate well using only direct translation, but when pivot translation

is used, their performance improves. In particular, the back-translation pivot baseline achieves slightly higher BLEU scores overall than any of the initial models trained without monolingual data.

Currey et al. (2017) showed that the copied corpus method was useful for adding target-language monolingual data to NMT training. Here, we see that the same method is not beneficial (and in fact is slightly harmful compared to the baseline) for adding pivot-language monolingual data to NMT. This could be because the copied corpus is used here to improve translation directions that are not of interest (i.e. translation into and out of English, rather than DE↔RU translation).

5.3 Proposed Models with Monolingual Data

We display the results for our three proposed models in the last three rows of Table 3. Compared to the best pivot-based model (back-translation), the pivot from scratch and pivot fine-tune models perform slightly better overall in both translation directions (DE→RU and RU→DE). Additionally, the pivot-parallel combined model improves over the best pivot-based model by about 1 BLEU for DE→RU and also does slightly better for RU→DE. This BLEU gain is especially interesting since the proposed models do not require two-step inference, unlike the back-translation pivot-based model.

Comparing to the best direct translation model (the zero-resource model based on Firat et al., 2016) leads to similar conclusions. The pivot from scratch and pivot fine-tune methods do similarly to this baseline for DE→RU translation and improve over it by 1.3–3.2 BLEU for RU→DE translation. For the pivot-parallel combined model, the gains over the baseline for DE→RU are stronger than for the other two methods, and the gains for RU→DE are similar. Thus, we have shown that adding pivot-language monolingual data through these methods can strongly improve zero-resource NMT performance.

All three of our proposed models improve over a strong direct translation baseline and perform similarly to or better than a pivot-based translation baseline that uses EN monolingual data without requiring the two-step inference process necessary for pivot-based translation. The pivot from scratch and pivot fine-tune models give similar results, while the pivot-parallel combined method, which adds in the back-translated parallel corpus, yields the best BLEU scores out of all models across the board.

	DE→RU		RU→DE	
BLEU	**iter 1**	**iter 2**	**iter 1**	**iter 2**
from scratch	**23.0**	**23.0**	22.3	**22.7**
fine-tune	23.0	**23.3**	22.4	**22.8**
combined	**23.6**	22.7	**22.5**	21.2

Table 4: BLEU scores for the proposed models on the test set (newstest2015). We show BLEU scores for one and two iterations (iter 1 and iter 2).

6 Iterating the Proposed Models

Inspired by Hoang et al. (2018) and Niu et al. (2018), we study whether iterating the proposed models can improve translation performance. Starting from the trained models from section 5.3, we run a second iteration as follows:

1. Back-translate the same EN data using the new model to create a new BT-pivot corpus (as described in section 3.2).

2. For the pivot-parallel combined method, back-translate the EN side of the parallel data as well (following Firat et al., 2016).

3. Fine-tune the model or train the model from scratch using the new data concatenated with the subset of the original parallel data (as described in section 3.3).

Table 4 shows the performance on the test dataset (newstest2015) when a second iteration of back-translation and training is performed. For the pivot from scratch and pivot fine-tune methods, we see small gains (up to 0.4 BLEU) from running a second iteration. These small improvements help the pivot from scratch and pivot fine-tune methods catch up to the single-iteration version of the pivot-parallel combined method. On the other hand, running a second iteration is very costly in terms of training time, since it requires another back-translation step and another training step. For the pivot-parallel combined model, which was the best-performing model with one iteration, adding a second iteration damages performance in terms of BLEU score. This seems to match the results of Hoang et al. (2018) that indicate that there are diminishing returns as more iterations are added.

7 Conclusions

This paper introduced the task of zero-resource neural machine translation using pivot-language monolingual data. We introduced a way of generating a pseudo-parallel source↔target training corpus using the monolingual pivot-language corpus, and we showed three ways of leveraging this corpus to train a final source↔target NMT system. All three methods improved over strong baselines that used both direct source→target translation and pivot translation through EN; the pivot-parallel combined method was the most successful.

Our proposed paradigm has several benefits. First, it shows that monolingual data from a language other than the source and target languages can aid NMT performance, complementing literature on using source- and target-language monolingual data in NMT. Second, this paradigm is architecture-agnostic, so it would be easy to apply to architectures that improve upon the basic zero-shot and zero-resource models (e.g. Arivazhagan et al., 2019; Lu et al., 2018; Platanios et al., 2018). However, the methods we have proposed are not without limitations. First, using the pivot-language monolingual data might not work as well when the source and target languages are closely related; this might be a case where source and target monolingual data is more useful than pivot monolingual data. These models also tune a multilingual NMT system for translation in two directions only (source→target and target→source), so they would not be applicable in cases where a single massively multilingual NMT system (Aharoni et al., 2019) is required.

In the future, we hope to additionally study the use of source-language and target-language monolingual data in zero-resource NMT. We would also like to test our proposed zero-resource methods on other zero-shot NMT architectures and on other language pairs. We also think that data selection methods on the back-translated data (Niu et al., 2018) could be helpful, since zero-shot multilingual NMT models often generate translations in the wrong target language (Arivazhagan et al., 2019).

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198. Association for Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277. Association for Computational Linguistics.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce

Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the Second Workshop on Neural Machine Translation and Generation*, pages 18–24. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V 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.

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

Surafel M Lakew, 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*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations*.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation*, pages 84–92. Association for Computational Linguistics.

Xing Niu, Michael Denkowski, and Marine Carpuat. 2018. Bi-directional neural machine translation with synthetic parallel data. In *Proceedings of the Second Workshop on Neural Machine Translation and Generation*, pages 84–91. Association for Computational Linguistics.

Jaehong Park, Jongyoon Song, and Sungroh Yoon. 2017. Building a neural machine translation system using only synthetic parallel data. *arXiv preprint arXiv:1704.00253*.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 86–96. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1715–1725. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

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. Association for Computational Linguistics.