Zero-Resource Translation with Multi-Lingual Neural Machine Translation

Orhan Firat[⋆]
Middle East Technical University
orhan.firat@ceng.metu.edu.tr

Baskaran Sankaran
IBM T.J. Watson Research Center

Yaser Al-onaizan
IBM T.J. Watson Research Center

Fatos T. Yarman Vural
Middle East Technical University

Kyunghyun Cho
New York University

Abstract

In this paper, we propose a novel finetuning algorithm for the recently introduced multi-way, multilingual neural machine translate that enables zero-resource machine translation. When used together with novel many-to-one translation strategies, we empirically show that this finetuning algorithm allows the multi-way, multilingual model to translate a zero-resource language pair (1) as well as a single-pair neural translation model trained with up to 1M direct parallel sentences of the same language pair and (2) better than pivot-based translation strategy, while keeping only one additional copy of attention-related parameters.

1 Introduction

A recently introduced neural machine translation (Forcada and Ñeco, 1997; Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014) has proven to be a platform for new opportunities in machine translation research. Rather than word-level translation with language-specific preprocessing, neural machine translation has found to work well with statistically segmented subword sequences as well as sequences of characters (Chung et al., 2016; Luong and Manning, 2016; Sennrich et al., 2015b; Ling et al., 2015). Also, recent works show that neural machine translation provides a seamless way to incorporate multiple modalities other than natural language text in translation (Luong et al., 2015a; Caglayan et al., 2016). Furthermore, neural machine translation has been found to translate between multiple languages, achieving better translation quality by exploiting positive language transfer (Dong et al., 2015; Firat et al., 2016; Zoph and Knight, 2016).

In this paper, we conduct in-depth investigation into the recently proposed multi-way, multilingual neural machine translation (Firat et al., 2016). Specifically, we are interested in its potential for zero-resource machine translation, in which there does not exist any direct parallel examples between a target language pair. Zero-resource translation has been addressed by pivot-based translation in traditional machine translation research (Wu and Wang, 2007; Utiyama and Isahara, 2007; Habash and Hu, 2009), but we explore a way to use the multi-way, multilingual neural model to translate directly from a source to target language.

In doing so, we begin by studying different translation strategies available in the multi-way, multilingual model in Sec. 3–4. The strategies include a usual one-to-one translation as well as variants of many-to-one translation for multi-source translation (Zoph and Knight, 2016). We empirically show that the many-to-one strategies significantly outperform the one-to-one strategy.

We move on to zero-resource translation by first evaluating a vanilla multi-way, multilingual model on a zero-resource language pair, which revealed that the vanilla model cannot do zero-resource translation in Sec. 6.1. Based on the many-to-one strategies we proposed earlier, we design a novel finetun-

[⋆] Work carried out while the author was at IBM Research.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 268–277,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

ing strategy that does not require any direct parallel corpus between a target, zero-resource language pair in Sec. 5.2, which uses the idea of generating a pseudo-parallel corpus (Sennrich et al., 2015a). This strategy makes an additional copy of the attention mechanism and finetunes only this small set of parameters.

Large-scale experiments with Spanish, French and English show that the proposed finetuning strategy allows the multi-way, multilingual neural translation model to perform zero-resource translation as well as a single-pair neural translation model trained with up to 1M true parallel sentences. This result re-confirms the potential of the multi-way, multilingual model for low/zero-resource language translation, which was earlier argued by Firat et al. (2016).

2 Multi-Way, Multilingual Neural Machine Translation

Recently Firat et al. (2016) proposed an extension of attention-based neural machine translation (Bahdanau et al., 2015) that can handle multi-way, multilingual translation with a shared attention mechanism. This model was designed to handle multiple source and target languages. In this section, we briefly overview this multi-way, multilingual model. For more detailed exposition, we refer the reader to (Firat et al., 2016).

2.1 Model Description

The goal of multi-way, multilingual model is to build a neural translation model that can translate a source sentence given in one of N languages into one of M target languages. Thus to handle those N source and M target languages, the model consists of N encoders and M decoders. Unlike these language-specific encoders and decoders, only a single attention mechanism is shared across all $M \times N$ language pairs.

Encoder An encoder for the n-th source language reads a source sentence $X = (x_1, \ldots, x_{T_x})$ as a sequence of linguistic symbols and returns a set of context vectors $C^n = \left\{\mathbf{h}_1^n, \ldots, \mathbf{h}_{T_x}^n\right\}$. The encoder is usually implemented as a bidirectional recurrent network (Schuster and Paliwal, 1997), and each context vector $\mathbf{h}_t^n$ is a concatenation of the forward and reverse recurrent networks' hidden states at time t. Without loss of generality, we assume that the dimensionalities of the context vector for all source languages are all same.

Decoder and Attention Mechanism A decoder for the m-th target language is a conditional recurrent language model (Mikolov et al., 2010). At each time step t', it updates its hidden state by

$$\mathbf{z}_{t'}^m = \varphi^m(\mathbf{z}_{t'-1}^m, \tilde{y}_{t'-1}^m, \mathbf{c}_{t'}^m),$$

based on the previous hidden state $\mathbf{z}_{t'-1}^m$, previous target symbol $\tilde{y}_{t'-1}^m$ and the time-dependent context vector $\mathbf{c}_{t'}^m$. φ^m is a gated recurrent unit (GRU, (Cho et al., 2014)).

The time-dependent context vector is computed by the shared attention mechanism as a weighted sum of the context vectors from the encoder C^n:

$$\mathbf{c}_{t'}^m = \mathbf{U} \sum_{t=1}^{T_x} \alpha_{t,t'}^{m,n} \mathbf{h}_t^n + \mathbf{b}, \tag{1}$$

where

$$\alpha_{t,t'}^{m,n} \propto \exp\left(f_{\text{score}}(\mathbf{W}^n \mathbf{h}_t^n, \mathbf{W}^m \mathbf{z}_{t'-1}^m, \tilde{y}_{t'-1}^m)\right). \tag{2}$$

The scoring function f_{score} returns a scalar and is implemented as a feedforward neural network with a single hidden layer. For more variants of the attention mechanism for machine translation, see (Luong et al., 2015b).

The initial hidden state of the decoder is initialized as

$$\mathbf{z}_0^m = \phi_{\text{init}}^m(\mathbf{W}^n \mathbf{h}_t^n). \tag{3}$$

With the new hidden state $\mathbf{z}_{t'}^m$, the probability distribution over the next symbol is computed by

$$p(y_t = w | \tilde{y}_{<t}, X^n) \propto \exp(g_w^m(\mathbf{z}_t^m, \mathbf{c}_t^m, \mathbf{E}_y^m\left[\tilde{y}_{t-1}\right]), \tag{4}$$

where g_w^m is a decoder specific parametric function that returns the unnormalized probability for the next target symbol being w.

2.2 Learning

Training this multi-way, multilingual model does not require multi-way parallel corpora but only a set of bilingual corpora. For each bilingual pair, the conditional log-probability of a ground-truth translation given a source sentence is maximize by adjusting the relevant parameters following the gradient of the log-probability.

3 Translation Strategies

3.1 One-to-One Translation

In the original paper by Firat et al. (2016), only one translation strategy was evaluated, that is, *one-to-one translation*. This one-to-one strategy works on a source sentence given in one language by taking the encoder of that source language, the decoder of a target language and the shared attention mechanism. These three components are glued together as if they form a single-pair neural translation model and translates the source sentence into a target language.

We however notice that this is not the only translation strategy available with the multi-way, multilingual model. As we end up with multiple encoders, multiple decoders and a shared attention mechanism, this model naturally enables us to exploit a source sentence given in multiple languages, leading to a *many- to-one translation* strategy which was proposed recently by Zoph and Knight (2016) in the context of neural machine translation.

Unlike (Zoph and Knight, 2016), the multi-way, multilingual model is not trained with multi-way parallel corpora. This however does not necessarily imply that the model cannot be used in this way. In the remainder of this section, we propose two alternatives for doing multi-source translation with the multi-way, multilingual model, which eventually pave the way towards zero-resource translation.

3.2 Many-to-One Translation

In this section, we consider a case where a source sentence is given in two languages, X_1 and X_2. However, any of the approaches described below applies to more than two source languages trivially.

In this multi-way, multilingual model, multi-source translation can be thought of as averaging two separate translation paths. For instance, in the case of Es+Fr to En, we want to combine Es→En and Fr→En so as to get a better English translation. We notice that there are two points in the multi-way, multilingual model where this averaging may happen.

Early Average The first candidate is to averaging two translation paths when computing the time-dependent context vector (see Eq. (1).) At each time t in the decoder, we compute a time-dependent context vector for each source language, $\mathbf{c}_t^1$ and $\mathbf{c}_t^2$ respectively for the two source languages. In this early averaging strategy, we simply take the average of these two context vectors:

$$\mathbf{c}_t = \frac{\mathbf{c}_t^1 + \mathbf{c}_t^2}{2}. \tag{5}$$

Similarly, we initialize the decoder's hidden state to be the average of the initializers of the two encoders:

$$\mathbf{z}_0 = \frac{1}{2}\left(\phi_{\text{init}}(\phi^1_{\text{init}}(\mathbf{h}^1_{T_{x_1}})) + \phi_{\text{init}}(\phi^2_{\text{init}}(\mathbf{h}^2_{T_{x_1}}))\right), \tag{6}$$

where ϕ_{init} is the decoder's initializer (see Eq. (3).)

Late Average Alternatively, we can average those two translation paths (e.g., Es→En and Fr→En) at the output level. At each time t, each translation path computes the distribution over the target vocabulary, i.e., $p(y_t = w|y_{<t}, X_1)$ and $p(y_t = w|y_{<t}, X_2)$. We then average them to get the multi-source output distribution:

$$\begin{aligned} p(y_t = w|y_{<t}, X_1, X_2) = \\ \frac{1}{2}(p(y_t = w|y_{<t}, X_1) + p(y_t = w|y_{<t})). \end{aligned} \tag{7}$$

An advantage of this late averaging strategy over the early averaging one is that this can work even when those two translation paths were not from a single multilingual model. They can be two separately trained single-pair models. In fact, if X_1 and X_2 are same and the two translation paths are simply two different models trained on the same language pair–direction, this is equivalent to constructing an ensemble, which was found to greatly improve translation quality (Sutskever et al., 2014; Jean et al., 2015)

Early+Late Average The two strategies above can be further combined by late-averaging the output distributions from the early averaged model and the late averaged one. We empirically evaluate this early+late average strategy as well.

4 Experiments: Translation Strategies and Multi-Source Translation

Before continuing on with zero-resource machine translation, we first evaluate the translation strategies described in the previous section on multi-source translation, as these translation strategies form a basic foundation on which we extend the multi-way, multilingual model for zero-resource machine translation.

4.1 Settings

When evaluating the multi-source translation strategies, we use English, Spanish and French, and focus on a scenario where only En-Es and En-Fr parallel corpora are available.

4.1.1 Corpora

En-Es We combine the following corpora to form 34.71m parallel Es-En sentence pairs: UN (8.8m), Europarl-v7 (1.8m), news-commentary-v7 (150k), LDC2011T07-T12 (2.9m) and internal technical-domain data (21.7m).

En-Fr We combine the following corpora to form 65.77m parallel En-Fr sentence pairs: UN (9.7m), Europarl-v7 (1.9m), news-commentary-v7 (1.2m), LDC2011T07-T10 (1.6m), ReutersUN (4.5m), internal technical-domain data (23.5m) and Gigaword R2 (20.66m).

Evaluation Sets We use newstest-2012 and newstest-2013 from WMT as development and test sets, respectively.

Monolingual Corpora We do not use any additional monolingual corpus.

Preprocessing All the sentences are tokenized using the tokenizer script from Moses (Koehn et al., 2007). We then replace special tokens, such as numbers, dates and URL's with predefined markers, which will be replaced back with the original tokens *after* decoding. After using byte pair encoding (BPE, (Sennrich et al., 2015b)) to get subword symbols, we end up with 37k, 43k and 45k unique tokens for English, Spanish and French, respectively. For training, we only use sentence pairs in which both sentences are only up to 50 symbols long.

See Table 1 for the detailed statistics.

# Sents	Train	Dev†	Test‡
En-Es	34.71m	3003	3000
En-Fr	65.77m	3003	3000
En-Es-Fr	11.32m	3003	3000

Table 1: Data statistics. †: newstest-2012. ‡: newstest-2013

4.2 Models and Training

We start from the code made publicly available as a part of (Firat et al., 2016)[1]. We made two changes to the original code. First, we replaced the decoder with the conditional gated recurrent network with the attention mechanism as outlines in (Firat and Cho, 2016). Second, we feed a binary indicator vector of which encoder(s) the source sentence was processed by to the output layer of each decoder (g_w^m in Eq. (4)). Each dimension of the indicator vector corresponds to one source language, and in the case of multi-source translation, there may be more than one dimensions set to 1.

We train the following models: four single-pair models (Es↔En and Fr↔En) and one multi-way, multilingual model (Es,Fr,En↔Es,Fr,En). As proposed by Firat et al. (2016), we share one attention mechanism for the latter case.

Training We closely follow the setup from (Firat et al., 2016). Each symbol is represented as a 620-dimensional vector. Any recurrent layer, be it in the encoder or decoder, consists of 1000 gated recurrent units (GRU, (Cho et al., 2014)), and the attention mechanism has a hidden layer of 1200 $\tanh$ units (f_{score} in Eq. (2)). We use Adam (Kingma and Ba, 2015) to train a model, and the gradient at each update is computed using a minibatch of at most 80 sentence pairs. The gradient is clipped to have the norm of at most 1 (Pascanu et al., 2012). We early-stop any training using the T-B score on a development set[2].

4.3 One-to-One Translation

We first confirm that the multi-way, multilingual translation model indeed works as well as single-pair models on the translation paths that were considered during training, which was the major claim

[1] `https://github.com/nyu-dl/dl4mt-multi`

[2] T-B score is defined as $\frac{\text{TER}-\text{BLEU}}{2}$ which we found to be more stable than either TER or BLEU alone for the purpose of early-stopping (Zhao and Chen, 2009).

	Src	Trgt	Multi		Single	
			Dev	Test	Dev	Test
(a)	Es	En	30.73	28.32	29.74	27.48
(b)	Fr	En	26.93	27.93	26.00	27.21
(c)	En	Es	30.63	28.41	31.31	28.90
(d)	En	Fr	22.68	23.41	22.80	24.05

Table 2: One-to-one translation qualities using the multi-way, multilingual model and four separate single-pair models.

		Multi		Single	
		Dev	Test	Dev	Test
(a)	Early	31.89	31.35	–	–
(b)	Late	32.04	31.57	32.00	31.46
(c)	E+L	32.61	31.88	–	–

Table 3: Many-to-one quality (Es+Fr→En) using three translation strategies. Compared to Table 2 (a–b) we observe a significant improvement (up to 3+ BLEU), although the model was never trained in these many-to-one settings. The second column shows the quality by the ensemble of two separate single-pair models.

in (Firat et al., 2016). In Table 2, we present the results on four language pair-directions (Es↔En and Fr↔En).

It is clear that the multi-way, multilingual model indeed performs comparably on all the four cases with less parameters (due to the shared attention mechanism.) As observed earlier in (Firat et al., 2016), we also see that the multilingual model performs better when a target language is English.

4.4 Many-to-One Translation

We consider translating from a pair of source sentences in Spanish (Es) and French (Fr) to English (En). It is important to note that the multilingual model was *not* trained with any multi-way parallel corpus. Despite this, we observe that the early averaging strategy improves the translation quality (measured in BLEU) by 3 points in the case of the test set (compare Table 2 (a–b) and Table 3 (a).) We conjecture that this happens as training the multilingual model has implicitly encouraged the model to find a *common context vector* space across multiple source languages.

The late averaging strategy however outperforms the early averaging in both cases of multilingual model and a pair of single-pair models (see Table 3 (b)) albeit marginally. The best quality was observed when the early and late averaging strategies were combined at the output level, achieving up to +3.5 BLEU (compare Table 2 (a) and Table 3 (c).)

We emphasize again that there was *no* multi-way parallel corpus consisting of Spanish, French and English during training[3]. The result presented in this section shows that the multi-way, multilingual model can exploit multiple sources effectively without requiring any multi-way parallel corpus, and we will rely on this property together with the proposed many-to-one translation strategies in the later sections where we propose and investigate zero-resource translation.

5 Zero-Resource Translation Strategies

The network architecture of multi-way, multilingual model suggests the potential for translating between two languages *without* any direct parallel corpus available. In the setting considered in this paper (see Sec. 4.1,) these translation paths correspond to Es↔Fr, as only parallel corpora used for training were Es↔En and Fr↔En.

The most naive approach for translating along a zero-resource path is to simply treat it as any other path that was included as a part of training. This corresponds to the one-to-one strategy from Sec. 3.1. In our experiments, it however turned out that this naive approach does not work at all, as can be seen in Table 4 (a).

In this section, we investigate this potential of zero-resource translation with the multi-way, multilingual model in depth. More specifically, we propose a number of approaches that enable zero-resource translation without requiring any additional bilingual or multi-way corpora.

5.1 Pivot-based Translation

The first set of approaches exploits the fact that the target zero-resource translation path can be decomposed into a sequence of high-resource translation paths (Wu and Wang, 2007; Utiyama and Isahara, 2007; Habash and Hu, 2009). For instance, in our

[3] We do not assume the availability of annotation on multi-way parallel sentence pairs. It is likely that there will be some sentence (or a set of very close variants of a single sentence) translated into multiple languages (eg. Europarl). One may decide to introduce a mechanism for exploiting these (Zoph and Knight, 2016), or as we present here, it may not be necessary at all to do so.

case, Es→Fr can be decomposed into a sequence of Es→En and En→Fr. In other words, we translate a source sentence (Es) into a pivot language (En) and then translate the English translation into a target language (Fr), all within the same multi-way, multilingual model trained by using bilingual corpora.

One-to-One Translation The most basic approach here is to perform each translation path in the decomposed sequence independently from each other. This one-to-one approach introduces only a minimal computational complexity (the multiplicative factor of two.) We can further improve this one-to-one pivot-based translation by maintaining a set of k-best translations from the first stage (Es→En), but this increase the overall computational complexity by the factor of k, making it impractical in practice. We therefore focus only on the former approach of keeping the best pivot translation in this paper.

Many-to-One Translation With the multi-way, multilingual model considered in this paper, we can extend the naive one-to-one pivot-based strategy by replacing the second stage (En→Fr) to be many-to-one translation from Sec. 4.4 using both the original source language and the pivot language as a pair of source languages. We first translate the source sentence (Es) into English, and use both the original source sentence and the English translation (Es+En) to translate into the final target language (Fr).

Both approaches described and proposed above do not require any additional action on an already-trained multilingual model. They are simply different translation strategies specifically aimed at zero-resource translation.

5.2 Finetuning with Pseudo Parallel Corpus

The failure of the naive zero-resource translation earlier (see Table 4 (a)) suggests that the context vectors returned by the encoder are not compatible with the decoder, when the combination was not included during training. The good translation qualities of the translation paths included in training however imply that the representations learned by the encoders and decoders are good. Based on these two observations, we conjecture that all that is needed for a zero-resource translation path is a simple adjustment that makes the context vectors from the encoder to be compatible with the target decoder. Thus, we propose to adjust this zero-resource translation path however without any additional parallel corpus.

First, we generate a small set of *pseudo bilingual pairs* of sentences for the zero-resource language pair (Es→Fr) in interest. We randomly select N sentences pairs from a parallel corpus between the target language (Fr) and a pivot language (En) and translate the pivot side (En) into the source language (Es). Then, the pivot side is discarded, and we construct a *pseudo* parallel corpus consisting of sentence pairs of the source and target languages (Es-Fr).

We make a copy of the existing attention mechanism, to which we refer as *target-specific attention mechanism.* We then finetune only this target-specific attention mechanism while keeping all the other parameters of the encoder and decoder intact, using the generated pseudo parallel corpus. We do not update any other parameters in the encoder and decoder, because they are already well-trained (evidenced by high translation qualities in Table 2) and we want to avoid disrupting the well-captured structures underlying each language.

Once the model has been finetuned with the pseudo parallel corpus, we can use any of the translation strategies described earlier in Sec. 3 for the finetuned zero-resource translation path. We expect a similar gain by using many-to-one translation, which we empirically confirm in the next section.

6 Experiments: Zero-Resource Translation

6.1 Without Finetuning

6.1.1 Settings

We use the same multi-way, multilingual model trained earlier in Sec. 4.2 to evaluate the zero-resource translation strategies. We emphasize here that this model was trained only using Es-En and Fr-En *bilingual* parallel corpora without any Es-Fr parallel corpus.

We evaluate the proposed approaches to zero-resource translation with the same multi-way, multilingual model from Sec. 4.1. We specifically select the path from Spanish to French (Es→Fr) as a target zero-resource translation path.

	Pivot	Many-to-1	Dev	Test
(a)			< 1	< 1
(b)	$\surd$		20.64	20.4
(c)	$\surd$	Early	9.24	10.42
(d)	$\surd$	Late	18.22	19.14
(e)	$\surd$	E+L	13.29	14.56

Table 4: Zero-resource translation from Spanish (Es) to French (Fr) *without* finetuning, using multi-way, multilingual model. When pivot is $\surd$, English is used as a pivot language.

6.1.2 Result and Analysis

As mentioned earlier, we observed that the multi-way, multilingual model *cannot* directly translate between two languages when the translation path between those two languages was not included in training (Table 4 (a).) On the other hand, the model was able to translate decently with the pivot-based one-to-one translation strategy, as can be seen in Table 4 (b). Unsurprisingly, all the many-to-one strategies resulted in worse translation quality, which is due to the inclusion of the useless translation path (direct path between the zero-resource pair, Es-Fr). Another interesting trend we observe is the Early+Late averaging (Table 4 (e)) seems to perform worse than Late averaging (Table 4 (d)) alone, opposite of the results in Table 3 (b-c). We conjecture that, by simply averaging two model outputs (as in E+L), when one of them is drastically worse than the other, has the effect of pulling down the performance of final results. But early averaging can still recover from this deficiency, upto some extent, since the decoder output probability function g_w^m (Eq. (4).) is a smooth function not only using the averaged context vectors (Eq. (5).).

These results clearly indicate that the multi-way, multilingual model trained with only bilingual parallel corpora is not capable of direct zero-resource translation *as it is*.

6.2 Finetuning with a Pseudo Parallel Corpus

6.2.1 Settings

The proposed finetuning strategy raises a number of questions. First, it is unclear how many pseudo sentence pairs are needed to achieve a decent translation quality. Because the purpose of this finetuning stage is simply to adjust the shared attention mechanism so that it can properly bridge from the source-side encoder to the target-side decoder, we expect it to work with only a small amount of pseudo pairs. We validate this by creating pseudo corpora of different sizes–1k, 10k, 100k and 1m.

Second, we want to know how detrimental it is to use the generated pseudo sentence pairs compared to using true sentence pairs between the target language pair. In order to answer this question, we compiled a true multi-way parallel corpus by combining the subsets of UN (7.8m), Europarl-v7 (1.8m), OpenSubtitles-2013 (1m), news-commentary-v7 (174k), LDC2011T07 (335k) and news-crawl (310k), and use it to finetune the model[4]. This allows us to evaluate the effect of the pseudo and true parallel corpora on finetuning for zero-resource translation.

Lastly, we train single-pair models translating directly from Spanish to French by using the true parallel corpora. These models work as a baseline against which we compare the multi-way, multilingual models.

Training Unlike the usual training procedure described in Sec. 4.2, we compute the gradient for each update using 60 sentence pairs only, when finetuning the model with the multi-way parallel corpus (either pseudo or true.)

6.2.2 Result and Analysis

Table 5 summarizes all the result. The most important observation is that the proposed finetuning strategy with *pseudo*-parallel sentence pairs outperforms the pivot-based approach (using the early averaging strategy from Sec. 4.4) even when we used only 10k such pairs (compare (b) and (d).) As we increase the size of the pseudo-parallel corpus, we observe a clear improvement. Furthermore, these models perform comparably to or better than the single-pair model trained with 1M *true* parallel sentence pairs, *although they never saw a single true bilingual sentence pair* of Spanish and French (compare (a) and (d).)

Another interesting finding is that it is only beneficial to use true parallel pairs for finetuning the multi-way, mulitilingual models when there are enough of them (1m or more). When there are only a small number of true parallel sentence pairs, we

[4] See the last row of Table 1.

				Pseudo Parallel Corpus				True Parallel Corpus			
	Pivot	Many-to-1		1k	10k	100k	1m	1k	10k	100k	1m
(a)	Single-Pair Models		Dev	–	–	–	–	–	–	11.25	21.32
			Test	–	–	–	–	–	–	10.43	20.35
(b)	√	No Finetuning		Dev: 20.64, Test: 20.4				–			
(c)			Dev	0.28	10.16	15.61	17.59	0.1	8.45	16.2	20.59
			Test	0.47	10.14	15.41	17.61	0.12	8.18	15.8	19.97
(d)	√	Early	Dev	19.42	21.08	21.7	21.81	8.89	16.89	20.77	22.08
			Test	19.43	20.72	21.23	21.46	9.77	16.61	20.40	21.7
(e)	√	Early+	Dev	20.89	20.93	21.35	21.33	14.86	18.28	20.31	21.33
		Late	Test	20.5	20.71	21.06	21.19	15.42	17.95	20.16	20.9

Table 5: Zero-resource translation from Spanish (Es) to French (Fr) *with* finetuning. When pivot is √, English is used as a pivot language. Row (b) is from Table 4 (b).

even found using pseudo pairs to be more beneficial than true ones. This effective as more apparent, when the direct one-to-one translation of the zero-resource pair was considered (see (c) in Table 5.) This applies that the misalignment between the encoder and decoder can be largely fixed by using pseudo-parallel pairs only, and we conjecture that it is easier to learn from pseudo-parallel pairs as they better reflect the inductive bias of the trained model and as the pseudo- parallel corpus is expected to be more noisy, this may be an implicit regularization effect. When there is a large amount of true parallel sentence pairs available, however, our results indicate that it is better to exploit them.

Unlike we observed with the multi-source translation in Sec. 3.2, we were not able to see any improvement by further averaging the early-averaged and late-average decoding schemes (compare (d) and (e).) This may be explained by the fact that the context vectors computed when creating a pseudo source (e.g., En from Es when Es→Fr) already contains all the information about the pseudo source. It is simply enough to take those context vectors into account via the early averaging scheme.

These results clearly indicate and verify the potential of the multi-way, multilingual neural translation model in performing zero-resource machine translation. More specifically, it has been shown that the translation quality can be improved even without any direct parallel corpus available, and if there is a small amount of direct parallel pairs available, the quality may improve even further.

7 Conclusion: Implications and Limitations

Implications There are two main results in this paper. First, we showed that the multi-way, multilingual neural translation model by Firat et al. (2016) is able to exploit common, underlying structures across many languages in order to better translate when a source sentence is given in multiple languages. This confirms the usefulness of positive language transfer, which has been believed to be an important factor in human language learning (Odlin, 1989; Ringbom, 2007), in machine translation. Furthermore, our result significantly expands the applicability of multi-source translation (Zoph and Knight, 2016), as it does not assume the availability of multi-way parallel corpora for training and relies only on *bilingual* parallel corpora.

Second, the experiments on zero-resource translation revealed that it is not necessary to have a direct parallel corpus, or deep linguistic knowledge, between two languages in order to build a machine translation system. Importantly we observed that the proposed approach of zero-resource translation is better both in terms of translation quality and data efficiency than a more traditional pivot-based translation (Wu and Wang, 2007; Utiyama and Isahara, 2007). Considering that this is the first attempt at such zero-resource, or extremely low-resource, translation using neural machine translation, we expect a large progress in near future.

Limitations Despite the promising empirical results presented in this paper, there are a number of shortcomings that needs to addressed in follow-up research. First, our experiments have been done only with three European languages–Spanish, French and English. More investigation with a diverse set of languages needs to be done in order to make a more solid conclusion, such as was done in (Firat et al., 2016; Chung et al., 2016). Furthermore, the effect of varying sizes of available parallel corpora on the performance of zero-resource translation must be studied more in the future.

Second, although the proposed many-to-one translation is indeed generally applicable to any number of source languages, we have only tested a source sentence in two languages. We expect even higher improvement with more languages, but it must be tested thoroughly in the future.

Lastly, the proposed finetuning strategy requires the model to have an additional set of parameters relevant to the attention mechanism for a target, zero-resource pair. This implies that the number of parameters may grow linearly with respect to the number of target language pairs. We expect future research to address this issue by, for instance, mixing in the parallel corpora of high-resource language pairs during finetuning as well.

Acknowledgments

OF thanks Iulian Vlad Serban and Georgiana Dinu for insightful discussions. KC thanks the support by Facebook, Google (Google Faculty Award 2016) and NVidia (GPU Center of Excellence 2015-2016).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? *arXiv preprint arXiv:1605.09186*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv:1406.1078*.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *ACL*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. ACL.

Orhan Firat and Kyunghyun Cho. 2016. DL4MT-Tutorial: Conditional gated recurrent unit with attention mechanism.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL*.

Mikel L Forcada and Ramón P Ñeco. 1997. Recursive hetero-associative memories for translation. In *Biological and Artificial Computation: From Neuroscience to Technology*, pages 453–462. Springer.

Nizar Habash and Jun Hu. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 173–181. Association for Computational Linguistics.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for wmt'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal, September. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*, pages 1700–1709.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *The International Conference on Learning Representations (ICLR)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation. *arXiv:1511.04586*.

Minh-Thang Luong and Christopher D Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. *arXiv:1604.00788*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015a. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015b. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. *INTERSPEECH*, 2:3.

Terence Odlin. 1989. *Language Transfer*. Cambridge University Press. Cambridge Books Online.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2012. On the difficulty of training recurrent neural networks. *arXiv preprint arXiv:1211.5063*.

Håkan Ringbom. 2007. *Cross-linguistic similarity in foreign language learning*, volume 21.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45(11):2673–2681.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *HLT-NAACL*, pages 484–491.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Bing Zhao and Shengyuan Chen. 2009. A simplex armijo downhill algorithm for optimizing statistical machine translation decoding parameters. In *HLT-NAACL*, pages 21–24.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *NAACL*.