Adapting Multilingual NMT to Extremely Low Resource Languages FBK's Participation in the Basque-English Low-Resource MT Task, IWSLT 2018

Surafel M. Lakew[†+]*, Marcello Federico*[⋆]

[†]University of Trento, [+]Fondazione Bruno Kessler, Trento, Italy
[⋆]Amazon AI, East Palo Alto, CA 94303, USA
lastname@fbk.eu, [⋆]marcfede@amazon.com

Abstract

Multilingual neural machine translation (M-NMT) has recently shown to improve performance of machine translation of low-resource languages. Thanks to its implicit transfer-learning mechanism, the availability of a highly resourced language pair can be leveraged to learn useful representation for a lower resourced language. This work investigates how a low-resource translation task can be improved within a multilingual setting. First, we adapt a system trained on multiple language directions to a specific language pair. Then, we utilize the adapted model to apply an iterative training-inference scheme [1] using monolingual data. In the experimental setting, an extremely low-resourced Basque-English language pair (i.e., $\approx$ 5.6K in-domain training data) is our target translation task, where we considered a closely related French/Spanish-English parallel data to build the multilingual model. Experimental results from an *i*) in-domain and *ii*) an out-of-domain setting with additional training data, show improvements with our approach. We report a translation performance of 15.89 with the former and 23.99 BLEU with the latter on the official IWSLT 2018 Basque-English test set.

1. Introduction

The amount and diversity of model training data have been shown to affect the performance of Neural Machine Translation (NMT) system [2]. The direct relation between dataset size and performance of NMT [3], calls for alternative approaches to improve low-resource language translation.

Multilingual models that constitute more than one language pair has been shown to improve the translation performance of the low-resources language direction [4, 5]. In it's simplified and most effective setting, building an M-NMT system requires only an additional "language-flag" on the data level. Then, the attentional encoder-decoder based NMT model can be trained with the aggregation of several language pairs. The flag functions as a mechanism to trigger and direct the generation of target tokens in a specific target language. Thus, when the training set is constructed with the merge of several language directions, the latent transfer-learning across languages within the conventional NMT architecture showed to improve low-resourced language pairs. However, M-NMT training mechanism is biased towards generating the language pair with the largest portion of training data [1]. This bias will consequently limit the expected level of improvement in translating low-resource language pairs.

In this work, we propose a progressive adaptation of a multilingual model to a single language pair. We cast the adaptation stage in iterative training-inference operations that utilize monolingual data. Assuming, the availability of a low-resource language pair and a high resource related/language pairs data, we specifically explore the following two mechanisms:

- Adapting a multilingual model trained with several language directions to a specific low-resourced language pair, with the aim to avoid ambiguities at the time of inference.
- Then, applying an iterative training-inference using monolingual data of the low-resourced pair, with the aim to acquire a more cleaner pseudo-parallel corpus for the next adaptation stage.

In our experimental setting, we apply the above two mechanisms for improving the extremely low-resourced (ELR) Basque(EU)-English(EN) language pair. Then, with the experimental results and discussion we present our participation of the IWSLT-2018[1] shared task on Low Resource MT of TED[2] talks from Basque to English direction. We evaluated our approach with the *i*) ELR training condition in a constrained in-domain data, and *ii*) by adding an out-of-domain training data in addition to the in-domain. We train both models in a similar language setting (i.e., the additional/related language pairs are French/Spanish-English).

For comparing our approach, we train a bilingual (Basque-English) baseline and multilingual baseline model by adding more data from the related language pairs. More specifically, to build the M-NMT model Basque-French and

(*) Work conducted while this author was at FBK.

[1] https://sites.google.com/site/iwsltevaluation2018/TED-tasks
[2] https://wit3.fbk.eu/

Basque-Spanish with a similar ELR condition, and French-English and Spanish-English with the relatively high resource data size are added to the bilingual model. All models share common configurations at training and inference time, unless stated differently. Models are trained following [4], preprocessing and training procedures using the Transformer model [6].

In the following sections, we begin by introducing NMT (§2). Following, we review the related work in multilingual models and transfer-learning (§3). In Section 4, we describe our model training approach, followed by dataset and preprocessing, experimental settings, and baseline models (§5). Finally, we give further analysis on the experimental results in Section 6.

2. Neural Machine Translation

A standard state-of-the-art NMT system comprises an encoder, a decoder and an attention mechanism, which are all trained with maximum likelihood in an end-to-end fashion [7]. Although there are different variants of the encoder-attention-decoder based approach, Recurrent variants being the predominant until recently [8], this work utilizes the "Transformer" model [6]. The encoder is purposed to encode a source sentence into hidden state vectors, whereas the decoder uses the last representation of the encoder to predict symbols in the target language. In a broad sense, the attention mechanism improves the prediction process by deciding which portion of the source sentence to emphasize at a time [9]. Nevertheless, in the Transformer architecture, the application of attention spans to the representation of encoder latent and decoder latent space.

The Transformer architecture works by relying on a self-attention (*intra-attention*) mechanism, removing all the recurrent operations that are found in the RNN approach. In other words, the attention mechanism is repurposed to compute the latent space representation of both the encoder and the decoder sides. However, with the absence of recurrence, *positional-encoding* is added to the input and output embeddings. Similarly, as the time-step in a recurrent network, the positional information provides the Transformer network with the order of input and output sequences.

In our work, we use the absolute positional encoding, but very recently the use of the relative positional information has been shown to improve performance [10]. The model is organized as a stack of encoder-decoder networks that works in an auto-regressive way, using the previously generated symbol as input for the next prediction. Both the decoder and encoder can be composed of uniform layers, each built of sub-layers, i.e., a multi-head self-attention layer and a position wise feed-forward network (FFN) layer. The multi-head sub-layer enables the use of multiple attention functions with a similar cost of utilizing attention, while the FFN sub-layer is a fully connected network used to process the attention sublayers; as such, FFN applies two linear transformations on each position and a ReLU [6].

3. Related Works

3.1. Multilingual NMT

Prior to the introduction of a shared attention mechanism [11], early works in multilingual NMT utilizes separate encoder, decoder and an attention mechanism to support the translation of either many-to-one [12], or one-to-many [13] language directions. Moreover, Firat et al. [11] introduced a many-to-many system, however, relying on separate encoder-decoder setup. In a simplified yet delivering better performance [4] and [5] introduced a "language-flag" based approach that shares the attention mechanism and a single encoder-decoder networks to enable multilingual models. In this work, we follow the Johnson et al. [4] approach for prepending a language-specific flag at the source side of the training and inference examples.

3.2. Transfer Learning and Model Adaptation

Zoph et al., (2016) [14], proposed how transfer-learning between two NMT models can improve a low-resourced MT task. In their approach, a language pair with the relatively large amount of data is first utilized to train a parent model, then the encoder-decoder parameters are transferred to initialize a child model for a low-resourced language pair. After initializing, in the fine-tunning stage, the parameters of the child decoder network is fixed. The main motivation behind updating only the encoder parameters is that the decoder language across the parent-child models stays the same. Similarly, the parent-child approach has been extended to analyze the effect of using related languages on the source side of the encoder-decoder network [15].

In a related way to benefit the low-resource language from the high resourced pair [16] proposed an alternative transfer-learning approach built on a component that allows to share lexical and sentence level representations of multiple source language to a single target language. In a prior work, a multi-source approach where two or more encoders shares an attention mechanism has been suggested in [17], to address the ambiguities of translating a source token to a single target language. Unlike [18] where a single multilingual model is used for several language translations [19] showed how adapting the multilingual model on a specific language pair improves performance. Recently [20] explored the advantage of initializing a low-resource language pair training using a pre-trained multilingual model showing a significant improvement over baseline approaches.

4. Adaptation from a Multilingual Model

This work aims to exploit the transfer-learning across languages, however, instead of the parent-child strategy [14], we rely on using a multilingual model as in [4] that allows to abstract the representation of several languages in a single attentional encoder-decoder model. We hypothesis if data is received both for the low and high resourced language

pairs, training a single model with the concatenation of all the data and progressively fine tunning it with the low-resource (*target-task*) language pair can avoid possible ambiguities between languages at the time of inference.

First, we train a model with all the available language pairs (including the target-task). Second, we adapt the best performing model to the target-task language pair. Unlike the recently proposed approach [20], we adapt using the same target-task data that has been utilized for training the baseline multilingual model. The main reason behind this is that the target-task data is already received at time of training the multilingual model. Then, the (latest) adapted model is used to perform back-translation [21] in a target $\rightarrow$ source direction or an iterative dual-inference in a $source \leftrightarrow$ target directions [1]. However, both inference approaches are used to create a source side synthetic data, the dual-inference requires an available monolingual data both from the source and target language. More importantly, the fact that we adapt from the multilingual to a bi-directional model allows us to avoid the use of auxiliary models (i.e., a separate model trained in a target $\rightarrow$ source direction) to perform the inference operations. After the inference stage, we continue training the model by combining the target-task and the newly formed source (synthetic) $\rightarrow$ target parallel data, consequently creating a progressive adaptation stages.

In the experimental section, the adaptation and progressive update of the multilingual model to the single language pair (Basque-English) target-task are evaluated in two settings:

- `iELR`, an extremely low-resource language pair trained and evaluated using an in-domain parallel and monolingual data.
- `oELR`, an extension of the `iELR` training condition with an additional out-of-domain parallel and monolingual data, as described in Section 5

In the following Section, the details of the experimental setup are given for the two evaluation scenarios.

5. Experiments

5.1. Dataset

The experimental setting covers the Basque (EU), English (EN), French (FR), and Spanish (ES) languages. The `ELR` language pair (EU-EN) and the related language pairs (FR-EU/EN, and ES-EU/EN) are categorized into the in-domain and out-of-domain settings. The in-domain data are extracted from the publicly released shared task dataset, WIT[3] TED corpus [22]. Where as the the out-of-domain dataset is collected from the WMT evaluation campaign PaCo corpus [23, 24], Opus corpus [25], and the Open Data Euskadi Repository (OpenData)[3]. Monolingual datasets for the Eu-EN pair are extracted from the TED, Opus18, and OpenData sources and preprocessed by removing the overlapping segments with the parallel data. Note; EN is the only available in-domain monolingual data, whereas the rest is collected from the out-of-domain sources based on availability. Table 1 summarizes the source and data size of each language direction.

	TED	Opus16/18	PaCo	OpenData
`EU-EN`	5623	856314	130359	-
`EU-FR`	5815	689358	-	-
`EU-ES`	5546	840458	-	926203
`FR-EN`	287134	-	-	-
`ES-EN`	277093	-	-	-
`EU-Mono`	-	-	-	741254
`EN-Mono`	242831	503970	-	-

Table 1: *Languages and dataset size of the training set. TED represents the in-domain data, whereas the Opus from the 2016 and 2018 (excluding the FR-EN and ES-EN pairs), PaCO for the EU-EN pair, and OpenData for the EU-ES pair represent the out-of-domain pairs.*

For evaluating the target-task (EU-EN) a development set of 1140 segments and for reporting the official submission results, the 2018 test set constituting 1051 source side segments are used from the TED talks in-domain data.

5.2. Preprocessing

We first tokenize the raw data and remove sentences longer than 70 tokens. As in [4], we prepend a "language-flag" on the source side of the corpus for all multilingual models. The internal sub-word segmentation [26] provided by the Tensor2Tensor library[4] is used before each training and inference. Note that prepending the "language-flag" on the source side of the corpus is specific to the multilingual models. Following the recommendation in [27], the number of segmentation rules is set to $16K$ for the in-domain data and $32K$ for the out-of-domain data.

5.3. Experimental Settings

All systems are trained using the Transformer [6] model implementation in the Tensor2Tensor library. For all trainings, we use the Adam optimizer [28], with an initial learning rate constant of 2 and a dropout [29, 30] of 0.2. The learning rate is increased linearly in the early stages ($warmup\_training\_steps$=$16,000$) and afterward it is decreased with an inverse square root of the training step.

Considering the two training scenario (i.e., `iELR` and `oELR`), we utilize two model configurations; *i*) for the in-domain data a 512 embedding and hidden units dimension, and 6 layers of self-attention encoder-decoder network, and *ii*) for an out-of-domain scenario the dimension is set to 1024. The training batch size is of $4,096$ sub-word tokens.

[3] http://hltshare.fbk.eu/IWSLT2018/OpendataBasqueSpanish.tgz

[4] https://github.com/tensorflow/tensor2tensor/tree/v1.6.2/tensor2tensor

	Round	NMT	M-NMT	iELR	oELR
Eu-En	I	3.10	13.37	12.96	22.48
	II	-	-	**15.65**	22.72
	III	-	-	15.15	**23.14**

Table 2: *BLEU results on the dev*2018 *using the EU-EN single language pair* NMT *and the multilingual* M-NMT *baseline models, as compared to the in-domain* iELR *and the out-of-domain* oELR *adapted multilingual models from three training rounds. The bold highlight shows the best performing training rounds.*

At inference time, we employ a beam size of 4 and a batch size of 32.

Following [6], iELR experiments are run upto 100k training steps, whereas oELR experiments are run upto 400K steps, i.e., all models are observed to converge within these steps. The consecutive adaptation converged in a variable training steps, however, to make sure a convergence point is reached, all restarted experiments are run for additional 50K steps. Then, the best performing checkpoint on the dev set is used in the next training stage. All models are trained on a Tesla V100-pcie-16gb with a single GPU for iELR and 4 GPU's for oELR.

5.4. Baseline Models

Baseline: models are trained as a term of comparison in two settings, *i*) using only the available in-domain EU-EN data, refereed to as NMT, and *ii*) by adding the related language (EU-FR/ES and FR/ES-EN) in-domain data on the EU-EN target-task. The latter forms a multilingual (M-NMT) baseline model. The following section, discusses the results and the comparison between the baselines and the adapted model types.

6. Results and Discussion

The baseline models (NMT and M-NMT) compared to against the adapted multilingual (iELR and oELR) models are reported in Table 2. The single language pair model trained with the in-domain (≈5.6K) training data showed a performance of 3.10 BLEU. As we expected, the poor performance is directly related to the small amount of training data. In case of the M-NMT, we observed an improvement of +10.27 over the NMT with a performance of 13.37 BLEU. As discussed in Section 1, the transfer-learning across languages, that arise from the additional EU-ES/FR and FR/ES-EN in-domain language pairs highly contributed for the observed improvement. Moreover, the experiments with our suggestion have been run for two consecutive rounds.

In-Domain Setting. In the first adaptation stage, the iELR model showed no significant difference with the baseline M-NMT. However, the adaptation stage helps to narrow the translation direction to the target task and avoid possible ambiguities for the inference stage. In the second round the iELR model showed a +2.28 BLEU improvement over the M-NMT (13.37 BLEU) baseline. The improvement is expected for the reason that the model is trained with the additional pseudo-parallel corpus from the back-translation step of the EN in-domain monolingual data to the EU target. In the consecutive round, however, the model performance degrades after the back-translation stage. This is likely caused by poorly generated source side synthetic EU from the EN monolingual data. Thus, for the final evaluation we take the best performing model from the second training round.

	NMT	M-NMT	iELR	oELR
Basque-English	-	-	15.89	23.99

Table 3: *Official BLEU results of tst*2018 *evaluated using the in-domain* iELR *and the out-of-domain* oELR *best performing adapted multilingual models.*

Out-of-Domain Setting: oELR models are trained in a similar training strategy with iELR, except the availability of additional parallel and monolingual (both for EU and EN, see Table 1) data. The relatively higher amount of training data, contributed for the larger gain of the oELR model over the in-domain training condition. Compared to the baseline models (NMT and M-NMT), oELR showed the highest performance with 23.14 BLEU score at the third training round. Unlike the performance degradation observed in the iELR setting, the availability of monolingual data both for EU and EN benefits each training-inference stage. However, with only a 0.66 BLEU gain over the initial model after three rounds, we observed that the domain mismatch between parallel (EU-EN) and the monolingual data disadvantages the expected improvement using the training-inference approach.

In case of, the official evaluation campaign, this work focused on a primary submission using the oELR model and a contrastive submission using the in-domain iELR model. Table 3, shows the performance of the two models on *test-2018*.

An interesting aspect from the multilingual adaptation and the iterative training-inference stages is that improvements are observed within 6k-20k steps. Meaning, the continued training approach from the latest adapted model shows a faster convergence than training a model from scratch. Overall, our approach aimed at training a baseline multilingual model for a progressive adaptation to a target-task (i.e., EU-EN), and applying an iterative training-inference scheme using monolingual corpora showed to improve over the baseline model. Our results suggest that the progressive adaptation is critical when the target-task language pair has new additional data at each stage. The experimental findings have brought our attention for a further study on how to adapt a multilingual model and what type of monolingual data to utilize in the training-inference stages.

7. Conclusions

In this work, we showed how progressively adapting a multilingual model to an extremely low-resourced (EU-EN) language pair improves the translation performance, with an additional training-inference stage that utilizes monolingual data. To evaluate the approach, the experimental setting is carried out in an in-domain (iELR) and out-of-domain (oELR) scenarios. Results show a significant improvement over a single language pair model (NMT), as well as a 2.28 BLEU increase over the baseline M−NMT model in an in-domain setting. As future work, we will focus on improving the joint iterative training-inference and progressive adaptation stages.

8. Acknowledgements

This work has been partially supported by the EC-funded project ModernMT (H2020 grant agreement no. 645487). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

9. References

[1] S. M. Lakew, Q. F. Lotito, N. Matteo, T. Marco, and F. Marcello, "Improving zero-shot translation of low-resource languages," in *14th International Workshop on Spoken Language Translation*, 2017.

[2] P. Koehn and R. Knowles, "Six challenges for neural machine translation," *arXiv preprint arXiv:1706.03872*, 2017.

[3] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, "Neural versus phrase-based mt quality: An in-depth analysis on english-german and english-french," *Computer Speech & Language*, vol. 49, pp. 52–70, 2018.

[4] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *arXiv preprint arXiv:1611.04558*, 2016.

[5] T.-L. Ha, J. Niehues, and A. Waibel, "Toward multilingual neural machine translation with universal encoder and decoder," *arXiv preprint arXiv:1611.04798*, 2016.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[9] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[10] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[11] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," *arXiv preprint arXiv:1601.01073*, 2016.

[12] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-task learning for multiple language translation." in *ACL (1)*, 2015, pp. 1723–1732.

[13] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," *arXiv preprint arXiv:1511.06114*, 2015.

[14] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," *arXiv preprint arXiv:1604.02201*, 2016.

[15] T. Q. Nguyen and D. Chiang, "Transfer learning across low-resource, related languages for neural machine translation," *arXiv preprint arXiv:1708.09803*, 2017.

[16] J. Gu, H. Hassan, J. Devlin, and V. O. Li, "Universal neural machine translation for extremely low resource languages," *arXiv preprint arXiv:1802.05368*, 2018.

[17] B. Zoph and K. Knight, "Multi-source neural translation," *arXiv preprint arXiv:1601.00710*, 2016.

[18] S. M. Lakew, Q. F. Lotito, T. Marco, N. Matteo, and F. Marcello, "Fbks multilingual neural machine translation system for iwslt 2017," in *14th International Workshop on Spoken Language Translation (IWSLT 2017)*, 2017, pp. 35–41.

[19] N.-Q. Pham, M. Sperber, E. Salesky, T.-L. Ha, J. Niehues, and A. Waibel, "Kits multilingual neural machine translation systems for iwslt 2017." IWSLT, 2017.

[20] G. Neubig and J. Hu, "Rapid adaptation of neural machine translation to new languages," *arXiv preprint arXiv:1808.04189*, 2018.

[21] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *arXiv preprint arXiv:1511.06709*, 2015.

[22] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[23] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz, *et al.*, "Findings of the 2016 conference on machine translation," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, vol. 2, 2016, pp. 131–198.

[24] I. San Vicente, I. Manterola, *et al.*, "Paco2: A fully automated tool for gathering parallel corpora from the web." in *LREC*, 2012, pp. 1–6.

[25] J. Tiedemann, "Parallel data, tools and interfaces in opus." in *Lrec*, vol. 2012, 2012, pp. 2214–2218.

[26] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[27] M. Denkowski and G. Neubig, "Stronger baselines for trustable results in neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, 2017, pp. 18–27.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[30] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in neural information processing systems*, 2016, pp. 1019–1027.