Adapting Multilingual Neural Machine Translation to Unseen Languages

Surafel M. Lakew[†+], *Alina Karakanta*[†+], *Marcello Federico*[+], *Matteo Negri*[+], *Marco Turchi*[+]

[†]University of Trento, [+]Fondazione Bruno Kessler, Trento, Italy

[†]name.surname@unitn.it, [+]surname@fbk.eu

Abstract

Multilingual Neural Machine Translation (MNMT) for low-resource languages (LRL) can be enhanced by the presence of related high-resource languages (HRL), but the relatedness of HRL usually relies on predefined linguistic assumptions about language similarity. Recently, adapting MNMT to a LRL has shown to greatly improve performance. In this work, we explore the problem of adapting an MNMT model to an unseen LRL using data selection and model adaptation. In order to improve NMT for LRL, we employ perplexity to select HRL data that are most similar to the LRL on the basis of language distance. We extensively explore data selection in popular multilingual NMT settings, namely in (zero-shot) translation, and in adaptation from a multilingual pre-trained model, for both directions (LRL↔en). We further show that dynamic adaptation of the model's vocabulary results in a more favourable segmentation for the LRL in comparison with direct adaptation. Experiments show reductions in training time and significant performance gains over LRL baselines, even with *zero* LRL data (+13.0 BLEU), up to +17.0 BLEU for pre-trained multilingual model dynamic adaptation with related data selection. Our method outperforms current approaches, such as massively multilingual models and data augmentation, on four LRL.[1]

1. Introduction

Neural Machine Translation (NMT) has become prevalent in the past years, contributing to the flow of information across languages and facilitating communication around the world. However, NMT requires a large amount of "feature-label" aligned data for building high-quality and usable systems [1]. For the majority of the world's languages, these resources are not available. Not benefiting from high quality MT (as it is usually the case with HRL) means that people's access to different sources of information can be restricted.

Multilingual Neural Machine Translation (MNMT) owes its success to cross-lingual knowledge transfer [2], which has been particularly beneficial for languages lacking large parallel data [3]. Previous works document further improvements when using languages from the same family, however they all rely on predefined linguistic assumptions about language similarity. Another challenge for facilitating access to information through MNMT is that relevant LRL data might not be available at the time of training the initial seed model, or not available at all. In most real-life applications, new needs in terms of domains or language coverage arise continuously, making monolithic MNMT models susceptible to out-of-vocabulary words. Moreover, new relevant training data in several (related or not) languages might become available continuously. Taking advantage of relevant data for adaptation is crucial to the performance of the final models [4, 5].

Recently, building a large scale MNMT model was shown to be beneficial for LRL [6], even outperforming models specifically fine-tuned on the LRL data [7]. Another approach optimizes embeddings through character n-grams (i.e., soft decoupled encoding, SDE) [8]. A more recent data augmentation approach showed improvements over all the previous approaches by adapting the MNMT system using pseudo-bitext generated by converting the HRL to the LRL [9]. Overall, research efforts in MT for LRL have shown that pre-training a multilingual NMT model and efficiently utilizing the available data are crucial towards better translation quality.

In this paper, we investigate the usefulness of language similarity (distance between languages) as an indicator for selecting *which* and *how much* related HRL data can lead to the largest possible improvements. In analysing these aspects, we examine the potential of a pre-trained universal (MNMT) model at two stages; *i*) without having access to the test language data at training time (zero-shot translation), and *ii*) after adapting it to the LRL with selected data based on a language similarity criterion. We evaluate our hypothesis in the following proposed settings;

Data Selection: We compute the perplexity of a LRL language model on available HRL data, in order to choose HRL data that are most similar to the LRL. Perplexity is a well-established information-theoretic measure, also used for measuring distance between languages [10]. We evaluate the data selection technique in different scenarios; including a) language family, b) random, and c) our proposed perplexity-based selection criterion.

Training and Inference: First we examine the performance of the universal model in total absence of LRL data (*zero-shot*). The evaluation involves both translation directions (LRL↔en). To date, model evaluation [7, 11, 8, 9] for the en→LRL has not been investigated yet. This direction

[1]Scripts to replicate the experiments and pre-trained models: https://github.com/surafelml/adapt-mnmt

is the most challenging one because of the small amount of available target side data in the LRL and the morphological richness of several LRL compared to English.

Adaptation of Pre-Trained Model: We experiment with the adaptation of the multilingual NMT system by preserving the initial model vocabulary (*DirAdapt*) or dynamically updating it to include new items (*DynAdapt*), as in [12]. Following previous observations that more frequent segmentation favors morphologically rich languages and LRL [13, 14], we extend this approach by choosing different segmentation sizes that improve performance on the LRL.

Based on the above three aspects, this work aims at finding a viable way to improve a LRL translation task. The contributions of our work are three-fold. In particular, we:

- Propose an effective data selection method to select relevant data from several related HRL that, on the same test languages, achieves better performance compared to the most recent data augmentation approach.
- Explore the extreme case of a total absence of training data in the test language by attempting zero-shot translation using a model trained with different portions of related HRL data in both translation directions.
- Explore and compare approaches that aim to improve the quality of LRL translation, including direct and dynamic adaptation of pre-trained models.

For a fair comparison with related works we utilize a standard dataset (TED Talks [15]) comprising 58 languages paired with English. Four languages are used as test. Two of them are extremely low-resource (Azerbaijani and Belarusian), while the other two (Galician and Slovak) are "relatively" low-resourced. We conduct our experiments using Transformer [16], which was shown to be superior for multilingual models [17] and in HRL benchmarks [18]. Experimental results show the effectiveness of our approach, which outperforms those presented in previous works.

2. Adapting Multilingual NMT

In this work, we aim to find a transfer-learning approach that leads to an efficient utilization of a pre-trained large-scale MNMT model. To achieve our goal of improving translation for the target LRL, we cast our approach as an unsupervised model adaptation strategy, in which relevant data for the adaptation are not supplied beforehand but have to be identified on the fly.

2.1. Data Selection by Language Distance

Perplexity is a commonly used measure to assess the quality of a language model [19], and has also been used to measure distance between languages [10].[2] In this work, we use perplexity to select HRL data that are similar to the LRL data. We train a language model on the LRL data (LRL_{LM}) and select training data with the lowest perplexity from related HRL (**Select-pplx**). We compare this approach with:

2. **Select-one** – Taking all available data only from one HRL related to the LRL.
3. **Select-fam** – Taking all available data from a set of HRL related to the LRL belonging to the same language family.
4. **Select-rand** – Randomly sampling an equal proportion of data with Select-one and Select-pplx, from the HRLs that are closely related to the LRL.

Perplexity is defined as the inverse probability of a test set (i.e., the HRL training data) computed using the LRL_{LM}. Thus, given the segments of the HRL set and the LM, the perplexity is computed as:

$$PP(S, LRL_{LM}) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1^{i-1})}} \quad (1)$$

Where: S is a HRL segment consisting the sequence $w_1, w_2, \ldots w_N$, P($\cdot$) are the n-gram probabilities estimated on the training set of LRL$_{LM}$. The distance between the LRL and the HRL is computed by evaluating the n-gram of the latter using the n-gram model of the former. For each HRL set, consisting of examples S_j, where $j = 1 \ldots m$, we select S_j with the lowest perplexity (i.e., closest to the LRL) by computing $PP(S_j, LRL_{LM})$. We repeat the process for each HRL, re-score the sentences of all HRL based on their perplexity and select the necessary portion of data determined by a pre-configured threshold.

2.2. Direct vs. Dynamic Adaptation

For adaptation, we pre-process the test language data either *i*) using the pre-trained model's segmentation rules, or *ii*) by first learning a new segmentation model from the LRL data. Thus, for the transfer-learning stage, we follow two strategies:

1. **DirAdapt**: Vocabularies, segmentation rules and all parameters of the pre-trained model are used without any change.
2. **DynAdapt**: New vocabularies are generated using the new segmentation rule, and portions of the pre-trained model parameter are re-used.

In the *DirAdapt* case, the segmentation rules of the pre-trained model are applied on the test language for the inference or adaptation stages. In the *DynAdapt* case, rules are

[2] We propose perplexity over popular data selection techniques in domain adaptation [4, 20], because the large number of languages involved makes training pairwise language models unfeasible for the scope of this work.

gl	size	Lang	Select-fam	Select-pplx
Train	10k	pt	184k	98.65k
Dev	682	es	196k	79.51k
Test	1,007	it	204k	6.85k
		Total:	584k	184k

Table 1: Data size for LRL gl, and selections using *Select-fam*, and *Select-pplx*.

learned from the test language data and new vocabulary items are generated accordingly. At adaptation time, if the entries in the test language vocabulary are already present in the current dictionary, all the relative pre-trained model weights are transferred, while a random initialization of the embedding layers and the pre-softmax linear transformation weight matrix is performed for newly inserted vocabulary items. Unlike [12], we first look for the test language segmentation that maximizes the overlap with the pre-trained model vocabularies.[3]

2.3. Zero-shot Translation

We specifically aim at assessing the potential of the large scale MNMT model towards zero-shot translation (ZST). Unlike with adaptation strategies, the translation is evaluated in an extreme scenario, where the LRL has never been seen at training time. This means that the transfer-learning to assist the LRL translation is expected to come from multiple languages, particularly related languages, that are present in the pre-trained model. We examine both a $LRL_{unseen} \leftrightarrow HRL$ translation directions, where:

1. $LRL_{unseen} \rightarrow$ HRL: represents a condition where the source side only sees related languages to the LRL, at training time but no LRL data at all.

2. HRL $\rightarrow LRL_{unseen}$: represents a so-far unexplored and more challenging condition, as discussed in Section 1.

To evaluate the two scenarios we pre-train several models with data featuring different size and language combinations. For constructing the data, we follow the perplexity-based data-selection criterion described in §2.1. In this proposal our objectives are; *i)* to evaluate how pre-trained models perform before an adaptation stage on unseen test language data, and *ii)* how models trained on data with different levels of language relatedness behave in addressing a zero-shot translation.

Our expectation is that the more closely related language pairs (HRL) to the test language (LRL_{unseen}) are available, the higher the performance of the pre-trained models will be. Comparing the zero-shot translation against the adapted models using a similar data selection criterion and data of the LRL_{unseen} will shade more light on how much the pre-training helps. Moreover, the zero-shot translation can signal how robustly both the encoder and the decoder learn without seeing the test language, but different combinations of related languages.

[3] An alternative approach could be to remove the embedding and projection layers of the pre-trained model, however, preliminary results showed lower performance; thus avoided from the scope of this work.

3. Experiments

3.1. Data and Preprocessing

For our experiments we use the TED talks corpus [15], which contains parallel data for 58 languages aligned to English. As a first step, we use four LRLs paired with English (en) for evaluating the two adaptation strategies; including Azerbaijan (az), Belarusian (be), Galician (gl), and Slovak (sk), and Turkish (tr), Russian (ru), Portuguese (pt), and Czech (cs) as their HRL respectively. All languages are used are used to train the massive MNMT model, except for the language serving as test language at each time. The choice of the test languages facilitates comparisons with previous works on similar settings. As a second step, we select a single test language pair (en↔gl) for an in-depth analysis of the data selection strategies, the zero-shot inference and the adaptation approaches. The data set size of the four LRL with their linguistically closest HRL are used as in [7].

Before each experiment, data is segmented into subword units using SentencePiece[4]. We use the same pre-processing both for NMT and LM experiments. Following the recommendation in [21], the segmentation rules are set to $8k$. The segmentation rules of the pre-trained models are used for the ZST and experiments using *DirAdapt*. Unless otherwise specified, the same number of segmentation rules is used for the *DynAdapt*, first by learning the rules using the test language.

3.2. Measuring Language Distance

For the related language data selection method, we focus on one language, Galician (gl) as the test language, paired with Portuguese (pt), in addition to Spanish (es) and Italian (it) as further auxiliary languages. First, we select pt as the closest language to gl (*Select-one*). Then, we include pt+es and pt+es+it for the experiments with selection based on the language family (*Select-fam*). For *Select-pplx*, we train a neural language model[5] on the Galician data to re-score sentences from the training corpora of related languages and select the sentences with the lowest perplexity until we match the corpus size of Portuguese. Selection is made without replacement, i.e., an English sentence can have translations in multiple languages. Statistics are shown in Table 1.

As proposed in Section §2.1, the distance between the test language to the rest of the languages is evaluated using perplexity of the Galician LM against the test sets of all other

[4] https://github.com/google/sentencepiece

[5] https://github.com/lverwimp/tf-lm

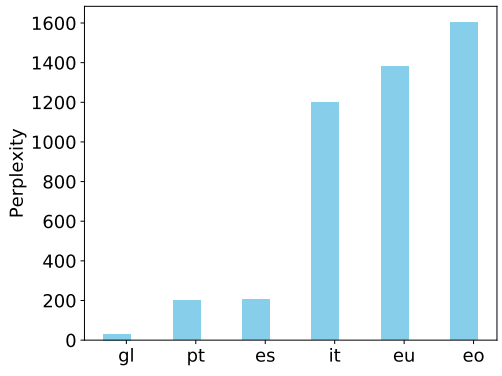

Figure 1: Perplexity for different languages (Portuguese/pt, Spanish/es, Italian/it, Basque/eu, Esperanto/eo) using the Galician/gl LM.

languages. Figure 1 shows the closest languages. The ranking reflects the proportion of each language in the mixed corpus with our perplexity-based selection. Even though Galician is considered to be more closely related to Portuguese, Spanish is behind it by only 4 perplexity points. Therefore, Spanish data is equally valuable for enhancing NMT performance on Galician.

In order to check if the improvement observed by adding more languages is due to simply having more training data, we set a maximum limit at time of data selection. We hence select the same amount of data (i.e., $184k$ for the case of Galician) using each of the following approaches: Select-one, Select-rand, and Select-pplx.

3.3. Model and Settings

The LM used for data selection is a 1-layer LSTM model with embedding and hidden layer of 512 units. We found that the best results were obtained when keeping all other settings as proposed in [22] (small model). We train the translation models with the OpenNMT[6] TensorFlow implementation of the Transformer model [23]. The model parameters are set to a 512 hidden unit and embedding dimension, 4 layers of self-attentional encoder-decoder with 8 heads. At training time, we use 4096 token level batch size with a maximum sentence length of 100. For inference, we keep a 32 example level batch size, with a beam search width of 5. LazyAdam [24] is applied throughout all strategies with an initial learning rate constant of 2. The learning rate increases linearly up to $8,000$ warm-up training steps, and decreases afterwards with an inverse square root of the training step. Given the sparsity of the test language data, dropout [25] is set to 0.3. The pre-trained models are run for up to 1M steps, and the adaptations steps vary based on the amount of data used. In all runs, models are observed to converge.

[6]http://opennmt.net/

3.4. Baselines and Comparison

Single language pair models (baselines) are trained from scratch using only the test LRL data. First, results from the adaptation and data-selection strategies are compared with these baselines. Then, we compare against solutions previously proposed in literature, namely:

- A direct adaptation of the multilingual model to the LRL, *RapAdapt* [7] and *SDE* [8].
- A massive multilingual model trained including all the test LRL, avoiding adaptation (*Many* $\leftrightarrow$ *Many*) [6].
- A data-augmentation for LRL pair, followed by adaptation of a multilingual model (*Data-Augment*) [9].

In the first two cases (*RapAdapt, SDE*), a similar strategy to our *DirAdapt* is implemented using an RNN model. For a fair comparison with our Transformer-based approach, we take the relative improvement (Δ) between the single pair baselines and the dynamically adapted models. The second (Many$\leftrightarrow$Many) and third (*Data-Augment*) approaches utilize the Transformer model, allowing us to directly compare against the reported results. More interestingly, these comparisons bring together several approaches using the same four test languages, aiming at improving the quality of LRL translation. As a metric to evaluate translation quality, we use BLEU [26].

4. Results and Analysis

4.1. Adaptation Does Matter

In Table 2, we show the Δ between the baseline of [7] (*RapAdapt*) and the best performing adaptation approach (*SDE*), against the Δ between our baseline and our best performing approach (*DynAdapt*). Even with stronger baselines, our Δ is higher than in the previous approaches with $+2.77$ BLEU averaged over the four test languages. Note that the MNMT model refers to a training setting with all except for the test language (cold start). We also note that our *DirAdapt* outperformed the *RapAdapt* and *SDE* with a larger margin in all the test languages.

The authors in [6] argue that the better performance of *Many*$\leftrightarrow$*Many* over the *RapAdapt* and *SDE* is due to avoiding model over-fitting by including more languages on both the encoder and decoder sides. However, our adaptation strategies show better performance in all test cases, with a $+1.37$ (*DirAdapt*) and $+2.78$ (*DynAdapt*) average BLEU. In fact, the additional improvement from *DirAdapt* comes from curating the segmentation for the test language and partially transferring the MNMT model parameters.

By contrasting the performance of previous works against the *DynAdapt*, we learn that our method is superior to all, in average BLEU score. Specifically, when compared to the latest *Data-Augment* [9], the *DynAdapt* shows better performance in two of the test languages (gl, sk), and slight

	Strategy	az[tr]	be[ru]	gl[pt]	sk[cs]	AVG.
Neubig & Hu 2018	Baseline	2.70	2.80	16.20	24.00	11.43
	MNMT→Bi (RapAdapt)	10.70	17.40	28.40	28.00	21.20
Wang et al., 2018	MNMT→Bi (SDE)	11.82	18.71	30.30	28.77	22.40
	Δ (SDE-Baseline)	9.12	15.91	14.10	4.77	10.98
Aharoni et al., 2019	Many ↔ Many	12.78	21.73	30.65	29.54	23.67
Xia et al., 2019	Data-Augment	**15.74**	**24.51**	33.16	32.07	26.37
Ours	Baseline	3.61	4.42	16.32	26.44	12.70
	MNMT→Bi (DirAdapt)	14.43	22.06	33.53	30.13	25.04
	MNMT→Bi (DynAdapt)	15.33	23.80	**34.18**	**32.48**	**26.45**
	Δ (DynAdapt-Baseline)	**11.72**	**19.38**	**17.86**	**6.04**	**13.75**

Table 2: BLEU scores for the four LRL→en comparing against previous approaches; RapAdapt [7], SDE [8], Many↔Many [6], and Data-Augment [9]. Bi is an adaptation with the LRL + [closest-HRL] according to *Select-one* strategy.

degradation for az and be. Our observation for the lower performance is that the data augmentation results in much larger synthetic data, while our adaptation utilized only the original LRL data for each of the test languages and the closest related language pair (amounting to a max of $200k$ segments) as in [7]. Overall, our approach showed the possibility of achieving better performance when initializing from pre-trained MNMT parameters.

4.2. Zero-shot Translation

Comparing the approaches in [7] that used RNNs for evaluating the ZST settings against our results, we observe a large difference (see Table 3) that again attests the superiority of the Transformer model. The better performance is particularly true for the MNMT models that are trained using all the available data but the test language. Previous works have also shown similar findings for the Transformer model when it comes to zero-shot translation [17, 6]. Thus, it is important to emphasize that the multilingual model is the best suit for further investigation by applying the data-selection procedures with the two adaptation options.

	Strategy	az	be	gl	sk
Neubig & Hu	Select-one	3.80	2.50	8.60	5.40
	MNMT	3.70	3.50	15.50	7.30
Ours	Select-one	3.25	2.07	13.59	9.30
	MNMT	**11.06**	**10.97**	**27.28**	**20.57**

Table 3: Results for LRL→en ZST using model trained with a single pair *Select-one*, and all but the test LRL (*MNMT*).

4.3. Data Selection for Zero-shot translation

Table 4 shows results for ZST using various data-selection strategies. In the gl→en direction, adding more data from related languages improves performance but the improvement slows down as more languages are added. Even without any test language data, performance increases from 13.59 BLEU for training only with pt (*Select-one*) to 24 BLEU for pt+es (*Select-fam*), while with it (pt+es+it) increases further by 1.34 BLEU. The MNMT model scores higher, but only by 1.77 BLEU when compared to best *Select-fam* strategy. Here, it is important to emphasize that: *i*) the MNMT model is trained using over 5M segments except for the test (gl-en) pair, meaning that the performance of *Select-fam* shows the possibility to improve a ZST by having more related languages but less data, *ii*) while the amount of data is the same for *Select-one*, *Select-rand*, and *Select-pplx*, the latter shows better performance, indicating the importance of the data selection criteria using perplexity.

Opposite results are obtained in the en→gl direction. As expected, our second evaluation of ZST into an unseen language on the decoder side does not perform well and performance decreases as more languages are added (i.e. from pt+es to pt+es+it). However, selecting related-language data using *Select-pplx*, we observe a better performance among the data-selection criteria at 5.38 BLEU.

Overall, ZST performance when translating from an unseen source language (gl) into a seen target language (en) is better than the baseline (see Table 4), with more than 10.0 BLEU points. This gain highlights the importance of closely related languages for improving the performance on the LRL. However, the opposite direction, where we infer into unseen target language (gl), is a more challenging task that requires further investigation and the availability of at least monolingual data for the LRL.

4.4. Data Selection for Adaptation

Table 5 shows results for adapting a MNMT model with data selected using our proposed perplexity-based method, both in the direct and the dynamic adaptation scenario. As a

	Strategy	gl→en	en→gl
Our non ZST	Baseline	16.32	**11.83**
Ours ZST	Select-one	13.59	8.05
	Select-rand	14.69	4.09
	Select-pplx	15.55	5.38
	pt+es	24.17	4.61
	pt+es+it	25.51	4.17
	MNMT	**27.28**	**8.78**

Table 4: BLEU for ZST using models trained with different data-selection criteria. Pt+es and pt+es+it are the two varieties of the *Select-fam* method.

MNMT Adaptation	gl→en	en→gl
Strategy	Dir/Dyn- Adapt	Dir/Dyn- Adapt
→ *gl*	32.18 / -2.5	26.39 / -3.21
→Select-one + *gl*	33.53 / +0.65	26.45 / +0.28
→Select-rand + *gl*	32.61 / +0.75	25.94 / +0.06
→Select-pplx + *gl*	$^{\uparrow}$**34.15** / $^{\uparrow}$**+1.41**	$^{\uparrow}$**27.35** / $^{\uparrow}$**+0.59**
→pt+es+it + *gl*	33.38 / +2.14	26.40 / +1.14

Table 5: BLEU using models adapted from the MNMT in different data conditions. $^{\uparrow}$ indicates statistically significance using bootstrap re-sampling ($p < 0.05$) [27].

general rule, adaptation with selecting data from several languages improves over adapting only with the target language. One possible reason for this improvement is avoiding overfitting to the little data of the target language, as shown in [7]. However, perplexity-based data selection (*Select-pplx*) outperforms selecting only one related language (*Select-one*) for both translation directions. Moreover, we show that the improvement does not come only from mixing several related languages, since *Select-rand* hurts performance for both directions. Our method improves even over adapting with all data from the most related languages (*pt+es+it*), allowing for a faster adaptation.

The results show that perplexity can be a reliable measure for selecting smaller amounts of related-language data both in translation and adaptation from a MNMT model in order to obtain larger improvements, with reduced training time (see Figure 2). For data selection strategies (either with perplexity or random), better performance is achieved with faster convergence. This confirms that the data-selection and the adaptation strategy is the fastest way to build a usable and better performing system for an unseen language from a pre-trained model.

Comparing the results of *DirAdapt* and *DynAdapt*, *DynAdapt* shows consistent improvements when adaptation is performed with data from at least two languages (LRL+another language). This can be attributed to the fact that the *DirAdapt* has a complete overlap (100%) both for the source and target side vocabularies with the pre-trained

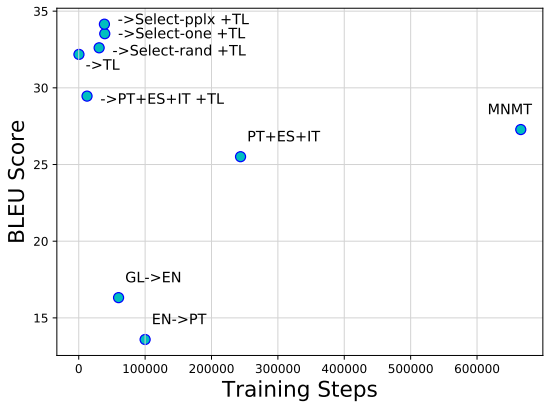

Figure 2: BLEU vs. training steps for gl→en direction.

initial model (i.e., the initial model vocabulary is used without any modification), as well as the transfer of all parameters when adapting. On the contrary, the *DynAdapt* improvement comes from a careful segmentation of the test language before adaptation, resulting in a new vocabulary and consequently enforcing a partial transfer of parameters from the initial model. In addition to the importance of data-selection, the additional gain using *DynAdapt* indicates that a universal multilingual model can be made stronger if tailored to the characteristics of the test languages when adapting.

When conducting a qualitative evaluation of the segmentation, for extremely low-resource test languages (such as gl↔en with $10k$ and az↔en with $5k$ bitext), we observed a frequent segmentation that favours sub-words closer to character level for most of rare words included in the vocabulary. This is consistent with previous work supporting character level segmentation for improving NMT of LRL [13, 14]. Furthermore, with a reduced vocabulary size, *DynAdapt* can compress the model with smaller embedding and pre-softmax linear transformation dimensions compared to the pre-trained model, and with sharing all the updated weight matrix as in [28].

5. Related Work

Multilingual NMT approaches share a common feature by aggregating data from various language pairs. In comparison with earlier approaches [29, 30, 31], training a single attentional encoder-decoder (universal) model using multiple pairs showed to be an efficient multilingual setting [3, 32]. While the performance of a universal model for HRL is comparable with a strong single language pair baseline, LRL pairs gain the highest improvement from the cross-lingual transfer. Thus, transfer learning for LRL can be defined in two main forms; *i*) "vertically", aggregating data from several language pairs to train a single model [3], *ii*) "horizontally", pre-training a model with the available pairs and fine-tuning it using the test (LRL) language data [33, 34, 12], or *iii*)with a combination of the two approaches.

Recently, new approaches have been introduced to efficiently adapt a pre-trained model to a LRL. One such case is proposed by [7], where they suggest to train a universal model (i.e., a model trained using up to 58 LRL-en pairs), with or without the test language direction. At time of adaptation, first, they adapt using only the LRL-en pair. Alternatively, a closely related language pair is added to the LRL-en as a regularizer when adapting. Both of the adaptation strategies show a larger performance gain over baseline models trained from scratch, however, their findings show the latter as an optimal adaptation setting.

Aimed at improving the source side language representation and parameter sharing, [8] introduced a multilingual lexicon encoding through character embedding, called Soft Decoupled Encoding. Their approach shows better performance than the adaptation strategies in [7], using a similar evaluation pairs. In a different work, a many-to-many multilingual model training is explored using all the available pairs both in LRL↔en directions [6]. By avoiding the adaptation stage, the approach showed to perform better when compared to the results in [7, 8] that utilize a many-to-one setting.

More recently, a data augmentation strategy is proposed to further improve the LRL pairs [9]. The approach leverages a target side monolingual and closely related HRL-English parallel data. Back translation is used to generate a pseudo-HRL from the monolingual data, while the HRL side of the parallel data is converted to pseudo-LRL using word substitution from a bilingual dictionary, similar to the approach in [35]. The synthetic data is used to construct a pseudo-HRL-en and a pseudo-LRL-en pair. Then, the synthetic data together with the available small LRL-en test language is used to improve over the baseline models. Using the same test languages, the data augmentation approach, which creates additional parallel data for the adaptation stage, outperformed the approaches reported in previous works [7, 8, 6].

This work shares a common ground on the effectiveness of pre-training a universal model and adaptating it to ultimately improve LRL pairs, however, it differs on the following aspects: *i*) it only considers a scenario where all of the pre-trained models have never seen the test language pair, *ii*) it learns a language model on the LRL to select data from related languages, *iii*) it investigates the less explored direction of en-LRL translation, *iv*) it explores zero-shot translation without adapting the pre-trained model, and *v*) it extends the dynamic adaptation strategy [12], that customizes any pre-trained model to the LRL pair.

In general, aggregating related HRL pair data with the LRL for an adaptation stage showed to perform better in all the test cases. Unlike in [7], who utilize only the immediately related language, we chose segments from different related languages based on the perplexity measure. Moreover, our approach does not rely on additional monolingual data or augmentation as in [9], instead, efficiently utilizes multiple related languages by identifying the relevant examples to the test language pair.

6. Conclusion

In this work, we focused on enhancing NMT performance for LRLs with data selection, and direct and dynamic adaptation of pre-trained models. To this aim, we used perplexity to select the most relevant data to the test language. We show that perplexity-based data selection improves translation, leading to an improvement of up to 10.0 BLEU points for LRL→en and 17.0 BLEU points for en→LRL when adapting from a multilingual model, with reduced training time. Our adaptation strategy with selected data is useful even in the extreme case of zero-shot translation for an unseen language (+13.0 BLEU). In future works, we plan to integrate our approach with data augmentation and semi-supervised model training strategies.

7. References

[1] P. Koehn and R. Knowles, "Six challenges for Neural Machine Translation," in *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39.

[2] O. Terence, "Language transfer-cross-linguistic influence in language learning," *Cambridge University Press. Cambridge Books Online.*, p. 222, June 1989.

[3] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation," *Transactions of the Association for Computational Linguistic*, vol. 5, pp. 339–351, 2017. [Online]. Available: https://aclweb.org/anthology/Q/Q17/Q17-1024.pdf

[4] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 355–362. [Online]. Available: http://aclweb.org/anthology/D11-1033

[5] M. v. d. Wee, A. Bisazza, and C. Monz, "Dynamic Data Selection for Neural Machine Translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 1400–1410. [Online]. Available: https://aclweb.org/anthology/D17-1147

[6] R. Aharoni, M. Johnson, and O. Firat, "Massively Multilingual Neural Machine Translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3874–3884.

[7] G. Neubig and J. Hu, "Rapid Adaptation of Neural Machine Translation to New Languages," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 875–880. [Online]. Available: https://aclweb.org/anthology/D18-1103

[8] X. Wang, H. Pham, P. Arthur, and G. Neubig, "Multilingual Neural Machine Translation With Soft Decoupled Encoding," in *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[9] M. Xia, X. Kong, A. Anastasopoulos, and G. Neubig, "Generalized Data Augmentation for Low-Resource Translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5786–5796.

[10] P. Gamallo, J. R. Pichel, and I. Alegria, "From language identification to language distance," *Physica A: Statistical Mechanics and its Applications*, vol. 484, pp. 152–162, 2017.

[11] J. Gu, H. Hassan, J. Devlin, and V. O. Li, "Universal neural machine translation for extremely low resource languages," in *Proceedings of NAACL-HLT 2018*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 344–354. [Online]. Available: https://aclweb.org/anthology/N18-1032

[12] S. M. Lakew, A. Erofeeva, M. Negri, M. Federico, and M. Turchi, "Transfer Learning in Multilingual Neural Machine Translation with Dynamic Vocabulary," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT), 2018*, 2018, pp. 54–61. [Online]. Available: https://workshop2018.iwslt.org/downloads/Proceedings_IWSLT_2018.pdf

[13] J. Kreutzer and A. Sokolov, "Learning to Segment Inputs for NMT Favors Character-Level Processing," in *Proceedings of the 15th International Workshop on Spoken Language Translation*, 2018, pp. 166–172.

[14] C. Cherry, G. Foster, A. Bapna, O. Firat, and W. Macherey, "Revisiting character-based neural machine translation with capacity and compression," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4295–4305. [Online]. Available: https://aclweb.org/anthology/D18-1461

[15] Y. Qi, D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig, "When and why are pre-trained word embeddings useful for neural machine translation?" in *Proceedings of NAACL-HLT 2018*. Association for Computational Linguistics, 2018, pp. 529–535. [Online]. Available: https://aclweb.org/anthology/N18-2084

[16] 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.

[17] S. M. Lakew, M. Cettolo, and M. Federico, "A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 641–652.

[18] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling Neural Machine Translation," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 1–9.

[19] R. Sennrich, "Perplexity minimization for translation model domain adaptation in statistical machine translation," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 539–549.

[20] R. C. Moore and W. Lewis, "Intelligent Selection of Language Model Training Data," in *Proceedings of the ACL 2010 Conference Short Papers*, 2010, pp. 220–224. [Online]. Available: https://www.aclweb.org/anthology/P/P10/P10-2041.pdf

[21] 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. [Online]. Available: https://www.aclweb.org/anthology/W17-3203

[22] L. Verwimp, J. Pelemans, H. V. Hamme, and P. Wambacq, "Character-Word LSTM Language Models," in *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 2017, pp. 417–427. [Online]. Available: https://arxiv.org/pdf/1704.02813.pdf

[23] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 67–72.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference for Learning Representations*, 2015.

[25] 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.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[27] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, vol. 4, 2004, pp. 388–395.

[28] O. Press and L. Wolf, "Using the Output Embedding to Improve Language Models," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 157–163.

[29] 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.

[30] 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.

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

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

[33] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, pp. 1568–1575. [Online]. Available: https://aclweb.org/anthology/D16-1163

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

[35] A. Karakanta, J. Dehdari, and J. van Genabith, "Neural machine translation for low-resource languages without parallel corpora," *Machine Translation*, vol. 32, no. 1, pp. 167–189, Jun 2018. [Online]. Available: https://doi.org/10.1007/s10590-017-9203-5