Multilingual Bidirectional Unsupervised Translation Through Multilingual Finetuning and Back-Translation

Bryan Li[1*], **Mohammad Sadegh Rasooli**[2], **Ajay Patel**[1], **Chris Callison-Burch**[1]
[1]University of Pennsylvania, Philadelphia, PA, USA
[2]Microsoft, Mountain View, CA, USA

Abstract

We propose a two-stage approach for training a single NMT model to translate unseen languages both to and from English. For the first stage, we initialize an encoder-decoder model to pretrained XLM-R and RoBERTa weights, then perform multilingual fine-tuning on parallel data in 40 languages to English. We find this model can generalize to zero-shot translations on unseen languages. For the second stage, we leverage this generalization ability to generate synthetic parallel data from monolingual datasets, then bidirectionally train with successive rounds of back-translation.

Our approach, which we EcXTra (English-centric Crosslingual (X) Transfer), is conceptually simple, only using a standard cross-entropy objective throughout. It is also data-driven, sequentially leveraging auxiliary parallel data and monolingual data. We evaluate unsupervised NMT results for 7 low-resource languages, and find that each round of back-translation training further refines bidirectional performance. Our final single EcXTra-trained model achieves competitive translation performance in all translation directions, notably establishing a new state-of-the-art for English-to-Kazakh (22.9 > 10.4 BLEU).

1 Introduction

Current neural machine translation (NMT) systems owe much of their success to efficient training over large corpora of parallel sentences, and consequently tend to struggle in low-resource scenarios and domains (Kim et al., 2020; Marchisio et al., 2020). This has motivated investigation into the field of zero-resource NMT, in which no parallel sentences are available for the source-target language pair. This is especially valuable for low-resource languages, which by nature have little to no parallel data.

There are two mainstream lines of inquiry towards developing models to tackle zero-resource machine translation. *Unsupervised machine translation* learns a model from monolingual data from the source and target languages. Some research involves introducing new unsupervised pre-training objectives between monolingual datasets (Lample and Conneau, 2019; Artetxe et al., 2019). Others devise training schemes with composite loss functions on various objectives (Ko et al., 2021; Garcia et al., 2021). In contrast, *zero-shot machine translation* learns a model by training on other datasets (Liu et al., 2020) or other language pairs (Chen et al., 2021, 2022), then directly employ this model for translating unseen languages.

This work leverages both mainstream approaches in zero-resource translation. We propose a conceptually simple, yet effective, two-stage approach for training a single NMT model to translate unseen languages both to and from English. The first stage model is trained on *real* parallel data from 40 high-resource languages to English. This results in a strong zero-shot model, which we use to translate unseen languages to English. By applying back-translation to flip the order, we obtain English-to-unseen *synthetic* parallel data. In the second stage, we continue training the model on successive rounds of offline back-translation, where each round uses the prior round for both for weight initialization and for synthetic parallel data.

We term our overall unsupervised translation approach EcXTra (English-centric Crosslingual (X) Transfer). EcXTra can be thought of as a data-driven approach, which sequentially leverages auxiliary parallel data then monolingual data. Each stage's model is initialized to an informed pretrained model, before fine-tuning. We initialize the first stage model's encoder and decoder to XLM-RoBERTa (Conneau et al., 2020) and RoBERTa (Liu et al., 2019) respectively, and we initialize the second stage model's weights to those

[*] Correspondence to: bryanli@seas.upenn.edu

Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023), pages 16–31
May 6, 2023 ©2023 Association for Computational Linguistics

of the first stage. In doing so, EcXTRa importantly avoids the complicated training schemes and custom training objectives of prior work.

As our approach is simple to train and extend to new unseen languages, we release all code, data and pretrained models.[1] Our contributions are:

1. We introduce EcXTra, a two-stage approach for training a single NMT model to translate unseen languages to and from English. In its two stages, EcXTra combines zero-shot NMT and unsupervised NMT: multilingual fine-tuning and back-translation respectively.

2. Our work is an empirical study of an agnostic view towards multilinguality, as we train the zero-shot stage on balanced splits of parallel data from 40 languages to English. In contrast, prior work has largely explored multilinguality by selecting train languages with oracle knowledge of the test languages.

3. We evaluate the bidirectional unsupervised NMT performance of a single EcXTra-trained model on 7 foreign-English test sets (14 total). This final model, trained in two rounds of back-translation, achieves competitive unsupervised performance for most language directions, establishing a new state-of-the-art for English-Kazakh. We are also the first to report, the best of our knowledge, unsupervised results for 3 translation directions: English-Pashto, English-Myanmar, and English-Icelandic.

2 Our Approach

Our training procedure closely follows the standard machine translation task. *Machine translation* involves developing models to output text in a target language $\mathcal{T}$, given text in a source language $\mathcal{S}$. In a typical supervised MT setting, it is assumed there is a parallel corpus $\mathcal{P} = \{(s_i, t_i)\}_{i=1}^{n}$ in which each sentence $t_i \in \mathcal{T}$ is a translation of $s_i \in \mathcal{S}$. A model is then trained on these examples, to minimize the cross-entropy loss given by

$$\mathcal{L}(\mathcal{P};\theta) = \sum_{i=1}^{n} \log p(t_i|s_i;\theta) \qquad (1)$$

where θ is a collection of learned parameters.

Given enough parallel data, this training framework allows contemporary NMT models to achieve strong performance (Dabre et al., 2020). However, in the unsupervised setting arises the fundamental challenge that we no longer have any parallel data between the source and target languages of interest.

Conceptually, we divide the two stages of our training procedure into four steps:

1a. *Zero-shot model transfer* by initializing to pretrained multilingual LMs. We use an XLM-RoBERTa encoder and a RoBERTa decoder.

1b. *Multilingual fine-tuning* for this initialized model, on parallel data from diverse source languages to English.

2a. *Synthetic parallel data creation* using back-translations from the stage 1 model.

2b. *Back-translation training* by initializing to the stage 1 model, then further training on the synthetic parallel data, in both translation directions. Steps 2a and 2b are iterated for several rounds, in each initializing to the prior round model.

Observe that these are are widely-used techniques in the field of machine translation. Our main contribution is in presenting an effective synthesis of the techniques to enable a single model to perform zero-shot and bidirectional translation (while using only a standard loss function).

Terminology It is worthwhile formalizing our exact terminology, given that prior work in this field uses terms rather inconsistently.[2] Our setting is *English-centric*, as the language pairs include English as either the source or target[3] Our final model is *bidirectional*, in that it can translate $\mathcal{S}$ to $\mathcal{T}$ and also translate $\mathcal{T}$ to $\mathcal{S}$. We call the non-English side of a pair a *foreign* language. Therefore, we use the terms foreign-English and many-to-English interchangeably (likewise with English-foreign and English-to-any). Languages seen during training on parallel datasets are *auxiliary* languages.

2.1 Zero-shot Model Transfer

There are many structural as well as lexical similarities across different languages, especially within language families. By training a multilingual translation model on gold-standard parallel datasets for auxiliary higher-resource languages, we aim to exploit these similarities. Specifically, we train model

[1] https://github.com/manestay/EcXTra

[2] See Section 2.1 of Garcia et al. (2021) for further discussion on this inconsistency.

[3] We focus on the English-centric setting because it is the language with the most parallel data to other languages.

parameters θ on parallel data between n auxiliary languages $\mathcal{S} = \mathcal{S}_1 \dots \mathcal{S}_n$ and some target language $\mathcal{T}$ (for us, English). The goal is to have the model learn to generalize to translating m unseen language data $\mathcal{U} = \mathcal{U}_1 \dots \mathcal{U}_m$ to $\mathcal{T}$. In other words, in the absence of gold-standard parallel data $\mathcal{P}$ in our zero-resource languages, we make use of knowledge transfer from larger parallel datasets with auxiliary source languages. Looking back at Equation 1, we redefine its objective function as

$$\sum_{i=1}^{n} \mathcal{L}(D(\mathcal{S}_i, \mathcal{T}); \theta) \tag{2}$$

where $D(\mathcal{S}_i, \mathcal{T})$ is the gold-standard parallel dataset for language $\mathcal{S}_i$ and English ($\mathcal{T}$).

EcXTRA: Multilingual fine-tuning Multilinguality, namely having diverse auxiliary languages is key to good zero-resource NMT performance (Garcia et al., 2021). In this setting, because there are no true (s_i, t_i) examples until inference time, performance becomes especially sensitive to the initialization of parameters θ. We do so by initializing the encoder with XLM-RoBERTa and decoder with RoBERTa. The former allows for transfer learning from strong pretrained models that are already trained on monolingual data in languages (including the unseen languages of interest), whereas the latter allows for a good understanding of fluent English sentences. Initializing the encoder and decoder to pretrained LMs follows prior work (Rothe et al., 2020; Ma et al., 2020).

From this initialization, we then fine-tune the model on parallel data from many high-resource languages to English. The resulting model is able to translate from unseen language to English, but not the other way. We next discuss how we extend our approach to develop a bidirectional model.

2.2 Synthetic Parallel Data Creation

We assume in this step that we have monolingual data in the unseen languages, which are typically collected by crawling web data. We make use of the model trained in the previous stage to translate all the monolingual sentences $(s_j)_{j=1}^{k}$ to English, thereby having synthetic parallel data $(s_j, \hat{t}_j)_{j=1}^{k}$ where $\hat{t}_i$ is the translation output from the zero-shot model. We then flip the order in each pair to produce examples $\hat{\mathcal{P}} = (\hat{t}_j, s_j)_{j=1}^{k}$, then continue training. This process of bootstrapping additional data is called (offline) *back-translation*.

While back-translation is typically used in low-resource settings, our approach extends it towards the zero-resource setting. We perform back-translation for all unseen languages, and concatenating together all synthetic parallel data $(\hat{\mathcal{P}}_i)_{i=1}^{m}$.

EcXTRA: Training on Synthetic Data In this step, we train a bidirectional English-centric model. We ensure bidirectionality by training on both the English-foreign synthetic parallel data, and the foreign-English auxiliary parallel data. Our new objective function is thus a combination of the two cross-entropy losses:

$$\sum_{i=1}^{n} \mathcal{L}(D(\mathcal{S}_i, \mathcal{T}); \theta) + \sum_{i=1}^{m} \mathcal{L}(\hat{\mathcal{P}}_i; \theta)$$

Just as we initialized the zero-shot model to pretrained multilingual LMs, so too do we initialize the unsupervised model to the zero-shot model. After training an initial unsupervised bidirectional model, we further refine performance by running iterative rounds of the synthetic parallel data creation and training process.

3 Datasets Used

Here we succinctly describe the data, providing further details in Appendix B.

Training For the zero-shot stage, we use parallel corpora from higher-resource auxiliary languages to English. We utilize a subset of the Many-to-English v1 dataset (Gowda et al., 2021). We consider only the 40 largest foreign-English pairs,[4] and equally sample 2 million examples from each.[5]

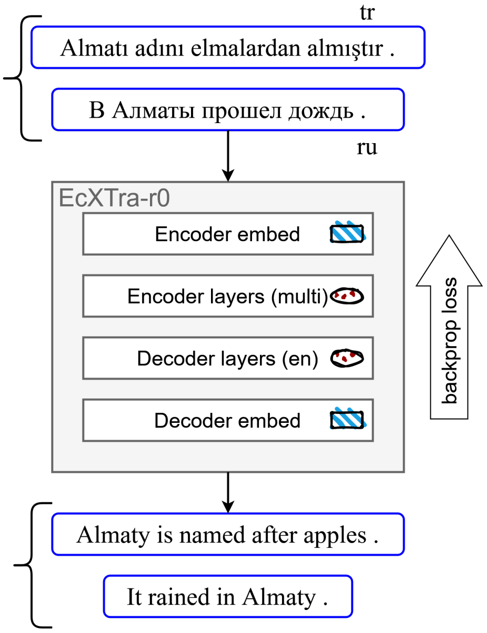
The resulting dataset, which we term *m2e-40*, consists of 80 million sentence pairs from 40 source languages. Note that unlike most prior work, we have taken an agnostic view towards multilinguality — we do not choose the training languages with reference to the testing languages.

For the unsupervised stage, we use monolingual corpora in the 7 test languages (below) from CommonCrawl and CC-100.

Testing We evaluate our approach on 7 languages: Kazakh (kk), Gujarati (gu), Sinhala (si), Nepali (ne), Pashto (ps), Icelandic (is), and Burmese (my). Test sets are taken from WMT21,

[4] Codes for training languages (with those used for validation in bold): **tr**, sr, **fr**, he, **ru**, ar, **zh**, bs, nl, **de**, pt, no, **it**, **es**, pl, **fi**, fa, sv, da, el, **hu**, sl, vi, **et**, sk, ja, **lt**, **lv**, uk, th, **cs**, ko, id, ca, mt, **ro**, bg, hr, **hi**, eu

[5] The rationale is further discussed in Section A.



Almatı adını elmalardan almıştır .

В Алматы прошел дождь .

Almaty is named after apples .

It rained in Almaty .

Figure 1: An illustration of the first stage of training (or EcXTra-r_0). The model learns to translate foreign sentences to English. The encoder is initialized to XLM-RoBERTa, and the decoder is initialized to RoBERTa. Both embeddings are frozen (blue rectangle), while layers are finetuned (red ellipse).

FLORES-101 and WAT21. The languages were chosen for both their diversity and for comparison to prior unsupervised NMT work.

Validation To validate the zero-shot stage, we select 15 foreign-English parallel datasets from WMT19 development data; these languages are seen during training.

In the unsupervised stage we only have access to monolingual data. For validation purposes, we thus reserve a small number of synthetic sentence pairs (250 per direction * 14 directions).

4 Experimental Setup

We move from the overall EcXTra approach, to the specifics of using EcXTra to train an NMT model.

4.1 Stage 1: Multilingual Fine-Tuning

Multilingual fine-tuning is the process of training a many-to-English zero-shot NMT model on parallel data from auxiliary languages to English. Figure 1 depicts the multilingual fine-tuning process.

Architecture We use an encoder-decoder, Transformer-based NMT model. Encoder layers and embeddings are initialized to XLM-R large, and decoder layers and embeddings are initialized to RoBERTa-large. These models were pretrained on a large multilingual corpora with various self-supervised language objectives. The encoder vocabulary is from XLM-R, and the decoder vocabulary is from RoBERTa.

Setup In the multilingual fine-tuning stage, we fine-tune our initialized model on WikiMatrix-25en. We freeze both the encoder and decoder embeddings and fine-tune both the encoder and decoder layers. This model thus has 0.76B trainable parameters (1.1B total). We select the best model checkpoint using early stopping.

Our training scheme uses the same supervised training objective of standard supervised NMT models. We hypothesize that this training scheme unlocks the cross-lingual transferability of XLM-R to zero-shot settings, with the same reasoning as Chen et al. (2022).

4.2 Stage 2: Back-Translation

In the unsupervised stage, we perform offline back-translation to bootstrap from foreign-English translation to English-foreign (and back). Figure 2 depicts the back-translation and training process.

Architecture Most of the architecture is transferred directly from the stage 1 model: encoder embeddings, encoder layers, and decoder layers. We cannot transfer the decoder embeddings, since the model now needs to output multiple languages. Instead, the decoder embeddings are tied to the encoder embeddings, which are frozen XLM-R embeddings. The resulting model thus has 0.96B trainable parameters (1.2B total parameters).

Notation Recall the zero-shot stage can be thought of as a pre-training step for the unsupervised stage. We thus designate the zero-shot model as EcXTra-r_0, and the unsupervised models as EcXTra-r_i, where i denotes the current round of back-translation (or simply r_i for brevity). We denote the *m2e-40* dataset as $\mathcal{D}_0$, the concatenation of all foreign monolingual corpora as $\mathcal{D}_{(l)}$, and the English monolingual corpus as $\mathcal{D}_{(e)}$. Synthetic parallel data are $\hat{\mathcal{D}}_{(l)\leftarrow(e)_i}$ or $\hat{\mathcal{D}}_{(e)\leftarrow(l)_i}$.

Training Data As 25M parallel sentences were used to train r_0, we generate about the same amount (3M per language * 8 languages = 24M) of back-translation data. Each r_i therefore is trained on ~50M sentences, given the bidirectional training.

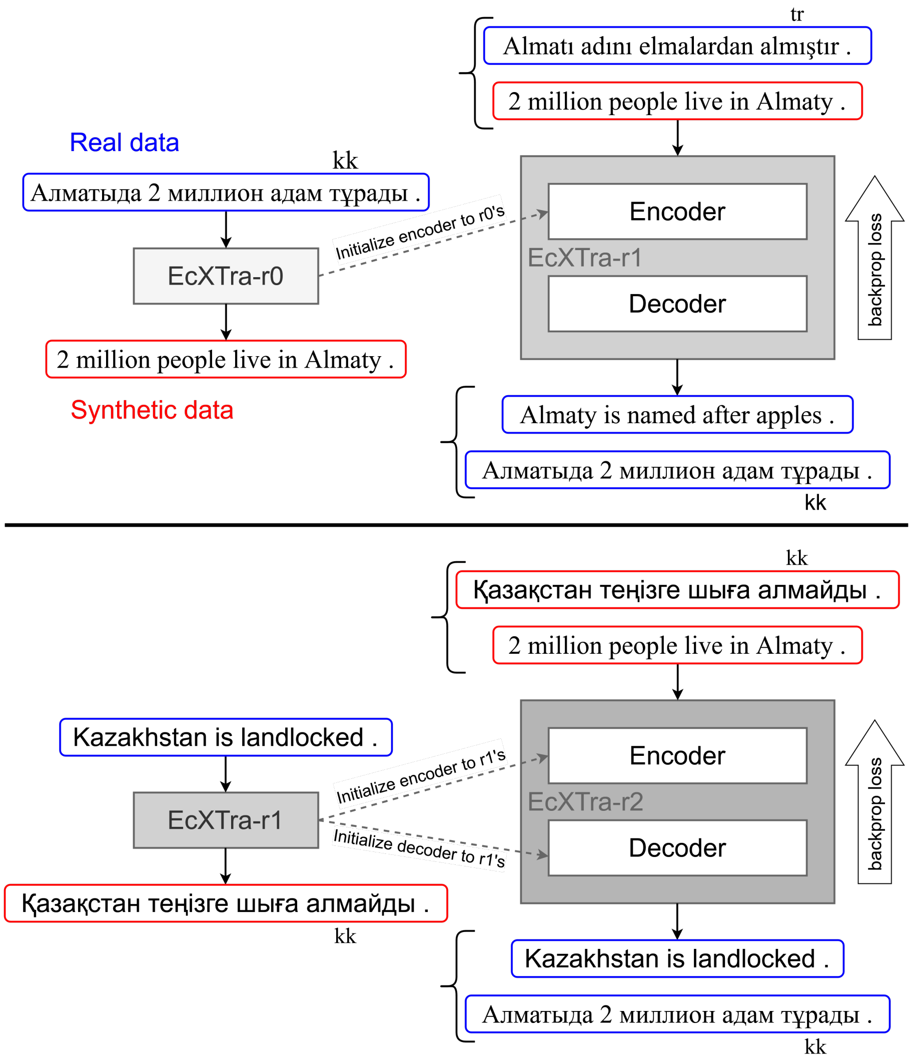

tr
Almatı adını elmalardan almıştır .
2 million people live in Almaty .
Real data
kk
Алматыда 2 миллион адам тұрады .
Initialize encoder to r0's
Encoder
EcXTra-r0
EcXTra-r1
Decoder
backprop loss
2 million people live in Almaty .
Synthetic data
Almaty is named after apples .
Алматыда 2 миллион адам тұрады .
kk
kk
Қазақстан теңізге шыға алмайды .
2 million people live in Almaty .
Kazakhstan is landlocked .
Initialize encoder to r1's
Encoder
EcXTra-r1
EcXTra-r2
Initialize decoder to r1's
Decoder
backprop loss
Қазақстан теңізге шыға алмайды .
kk
Kazakhstan is landlocked .
Алматыда 2 миллион адам тұрады .
kk

Figure 2: An illustration of the second stage of training, split into 2 rounds. Each round n is trained on a concatenation of back-translations from round $n-1$, and the the opposite direction training data from round $n-1$. Round 1 uses English-foreign synthetic data and transfers only the encoder, while round 2 uses synthetic data for both directions and transfers both encoder and decoder. Note that EcXTra blocks are abbreviated from Figure 1.

For each source language sentence, we add a special start token to indicate the desired target language, following the trick of Johnson et al. (2017). An example is <2kk> to target Kazakh.[6]

Setup Back-translation proceeds in successive stages. The main idea is that, for the current round r_i, we use r_{i-1} to generate synthetic parallel data by translating the monolingual corpus—$\mathcal{D}_{(l)}$ for odd rounds, $\mathcal{D}_{(e)}$ for even rounds. The source and target directions are then flipped before being used as training data. We also use r_{i-1} to intialize weights for r_i.

In our approach we aim to train bidirectional models. Therefore, the training data of r_i consists of both back-translations from r_{i-1}, as well as the opposite direction training data used for r_{i-1} itself. Thus the training data for round 1 is $\hat{\mathcal{D}}_{(l)\leftarrow(e)_1} + \mathcal{D}_0$, and for round 2 is $\hat{\mathcal{D}}_{(e)\leftarrow(l)_2} + \hat{\mathcal{D}}_{(l)\leftarrow(e)_1}$.

We ensure that for synthetic parallel data, the target side is always fluent monolingual text. As observed by Niu et al. (2018), this avoids the possible degradation from training to produce MT output.

For our experiments, we set $m = 2$, performing two rounds of back-translation – consistent with prior findings that improvement tapers off after two rounds (Hoang et al., 2018). The final model, EcXTra-r_2, will have learned from back-translated data in both directions.

[6] Our specific implementation is detailed in Appendix D.

Round	kk-en		gu-en		si-en		ne-en		ps-en		is-en		my-en		Avg.	
	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←
r_0	19.6	n/a	23.2	n/a	17.5	n/a	20.9	n/a	9.8	n/a	26.0	n/a	16.5	n/a	19.1	n/a
r_1	**18.5**	20.7	21.1	13.1	14.8	6.6	18.0	8.3	9.0	8.0	24.4	23.4	**14.3**	8.3	17.2	12.6
r_2	18.2	**22.9**	**21.5**	**13.9**	**17.8**	**7.1**	**19.7**	**9.3**	**13.0**	**8.1**	**30.6**	**25.4**	12.9	**8.8**	**19.1**	**13.6**

Table 1: BLEU scores for various rounds of EcXTra models on several low-resource translation test sets. The row divisions indicate groups by approach: zero-shot (no synthetic parallel data), unsupervised (synthetic parallel data). Foreign-English translation (→) columns are in white, while English-foreign (←) columns are in grey. 'Avg.' is the unweighted average BLEU scores across that translation direction. 'n/a' indicates unsupported directions. For the second group, the best BLEU score per column is **bolded**.

5 Results

We evaluate our models on test sets for 7 low-resource-to-English pairs in both translation directions (14 directions total). We use evaluation metrics with are consistent with prior work. By default, we report detokenized sacreBLEU (Post, 2018).[7] For the Indic languages (gu, si, ne), we report tokenized BLEU with the Indic-NLP library (Kunchukuttan, 2020). For Burmese (my), we report SPM-BLEU (Goyal et al., 2022) to handle the language's optional spacing.

5.1 Main Results

Table 1 shows results for each EcXTra round.

Foreign-English Results (→) EcXTra-r_0 (or r_0) is indeed able to perform zero-shot foreign-English translations. The unsupervised r_1 has lower scores, this is likely because this model is now tasked with performing 7 additional tasks on top of the original many-to-English task. r_2 recovers the overall performance, with the same average BLEU as r_0. While r_2 underperforms r_1 for a few individual pairs, it handily beats r_0 for ps-en (13.0 > 9.8) and for is-en (30.6 > 26.0), underscoring the overall quality of the back-translations.

English-Foreign Results (←) Similarly for English-foreign, we observe that r_2 matches or exceeds r_1 overall across language pairs (13.6 > 12.6). This is in spite of r_1 and r_2 sharing the same English-foreign training data $\mathcal{D}_{(l)\leftarrow(e)_1}$.

5.2 Comparisons with Prior Work

Table 2 compares the best EcXtra-trained model, r_2, with prior work (as well as the zero-shot r_0).[8] We emphasize that these results are *not fully comparable*, given the differing training datasets, models, and number of languages supported.[9] However, the comparisons can still illustrate the effectiveness of the language-agnostic nature and simplicity of EcXTra. We compare to:

SixT (Chen et al., 2021): trained on a German-English parallel dataset.

SixT+ (Chen et al., 2022): trained on AUX6, a parallel dataset in 6 high-resource languages. This is concurrent to our work.

mBART-ft (Tang et al., 2021): mBART-ft is an mBART model further fine-tuned on AUX6.

Garcia et al. (2021) : a single bidirectional unsupervised NMT model trained in 3 stages using combinations of various training objectives on parallel data, real and synthetic (from back-translation).

Zero-Shot NMT Results Considering the first four rows of Table 2 we see that EcXTra-r_0 outperforms mBART-ft and SixT for all translation pairs. Overall, it underperforms SixT+ (a concurrent work), but ties for si-en, and bests it for my-en (16.5 > 15.3).[10]

Unsupervised NMT Results We next compare our best unsupervised model, EcXTra-r_2 to Garcia et al. (2021), the only prior work, to the best of our knowledge, that also trains a single bidirectional unsupervised NMT model. r_2 notably achieves a new state-of-the-art for unsupervised en-kk (22.9 > 10.4), and also improves on kk-en (18.2 > 16.4) and si-en (17.8 > 16.2). r_2 underperforms for gu-en (13.9 < 16.4) and ne-en (19.7 < 21.7).

[7] `BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0`

[8] Confidence intervals for our results are not shown, but fall between ±0.4 to ±1.0.

[9] More discussion can be found in Section A.

[10] Chen et al. (2022) did not provide is-en results, but their model should support it.

Round	kk-en		gu-en		si-en		ne-en		ps-en		is-en		my-en	
	→	←	→	←	→	←	→	←	→	←	→	←	→	←
mBART-ft	19.6	n/a	17.3	n/a	12.2	n/a	14.4	n/a	0.9	n/a	...	n/a	3.6	n/a
SixT	19.0	n/a	17.3	n/a	12.2	n/a	14.4	n/a	11.4	n/a	...	n/a	5.4	n/a
SixT+	**27.3**	n/a	**27.5**	n/a	**17.5**	n/a	**23.7**	n/a	**12.9**	n/a	...	n/a	15.3	n/a
EcXTra-r_0	19.6	n/a	23.2	n/a	**17.5**	n/a	20.9	n/a	9.8	n/a	26.0	n/a	**16.5**	n/a
Garcia et al. (2021)	16.4	10.4	**22.2**	**16.4**	16.2	**7.9**	**21.7**	8.9	n/a	n/a	n/a	n/a	n/a	n/a
EcXTra-r_2	**18.2**	**22.9**	21.5	13.9	**17.8**	7.1	19.7	**9.3**	13.0	8.1	30.6	25.4	12.9	8.8
Supervised[1234567]	...	12.1	...	28.2	...	6.5	...	26.3	...	11.0	...	23.6	...	13.9

Table 2: BLEU scores comparing various models to EcXTra. The row divisions indicate groups by approach: zero-shot (no synthetic parallel data), unsupervised (synthetic parallel data), and supervised (real parallel data). 'n/a' indicates unsupported directions, while '...' indicates results not provided. Within a row group, the best BLEU score per column is **bolded**. Supervised results, from left to right: [1]Rasooli et al. (2021) [2]Li et al. (2019) [3]Bei et al. (2019) [4]Ko et al. (2021) [5]Shi et al. (2020) [6]Símonarson et al. (2021) [7]Hlaing et al. (2021)

Our work is the first to report unsupervised NMT on en-ps, en-is, and en-my. For an upper bound we cite prior results from supervised NMT systems; these are for reference only (and not even necessarily bidirectional nor multilingual). As expected, r_2 underperforms for most tasks. However, r_2 notably exceeds supervised results for en-is (25.4 > 23.6), showing the strength of our approach.

6 Discussion and Analysis

Enabling English-foreign translation in the second stage seems to come at the cost of some foreign-English performance. This may be an instance of the insufficient modeling capacity problem of multilingual NMT models (Zhang et al., 2020). Still, r_2 improves over r_1, while training on entirely synthetic parallel data generated from back-translations in both directions. This finding underscores the effectiveness of successive rounds of back-translation.

The EcXTra-trained model r_0 underperforms SixT+ (Chen et al., 2022) for foreign-English translations. Because EcXTra is a training approach, we can use SixT+ as a drop-in replacement for r_0 for both weight initialization, and for its back-translations. We suspect that training such a combined model would achieve even better English-foreign performance, and leave this to future work.

The EcXTra-trained model r_2 underpeforms Garcia et al. (2021) for English-Indic translations. This is likely a function of our *m2e-40* dataset having a much lower proportion of Hindi that the dataset of Garcia et al. (2021).[11] While we take an agnostic view of multilinguality, our training data is by no means writing script-centric; possibly making our model worse at outputting Indic texts. The exceeding en-kk and high en-is scores of r_2 provide some evidence for this.

Overall, the r_2 achieves competitive unsupervised translation results. Our model supports 3 additional language pairs over prior bidirectional unsupervised translation models, and the EcXTra approach makes it simple to extend to even more translation pairs. We underscore the overall appeal of our approach, in that we can use the zero-shot model to bootstrap back-translations for any unseen language, and train a bidirectional translation system from there.

6.1 Many-to-English Performance of Unsupervised Models

Unlike for the zero-shot r_0, the unsupervised r_2 has seen text in the text languages, albeit as synthetic parallel sentences with English. A natural question to ask is whether r_2 is able to maintain many-to-English performance for non-test languages.

We perform the following experiment to examine this. The models are tasked with *supervised* translation from 4 train languages (zh, hi, tr, ru) to English. r_0 and r_1 directly see these in their training parallel data, whereas r_2 has only indirectly seen them through the prior rounds.

The results are shown in Table 3. As was found for the test languages, r_1 performs worse than r_0. r_2 has the same average BLEU across language pairs as r_1. From this short experiment we have

[11] This is not explicitly specified in their paper, but is clear given their 4 auxiliary languages, vs our 40.

Round	zh-en	hi-en	tr-en	ru-en	Avg.
r_0	19.2	21.9	28.5	34.0	25.9
r_1	17.0	17.6	26.2	32.5	23.3
r_2	17.4	16.0	27.1	32.9	23.3

Table 3: BLEU scores for each EcXtra training round on several supervised foreign-English translations.

shown that the unsupervised models r_1 and r_2 do retain reasonable Many-to-English performance. We leave future work to investigate mitigation of the forgetting of prior learned tasks, endemic to (almost) all deep learning-based models.

7 Related Work

The field of low-resource and zero-resource neural machine translation is an area of continued interest. Below, we describe related works those which follow our data constraint: parallel foreign-English data in auxiliary languages, and monolingual data in unseen languages.

7.1 Many-to-English zero-shot NMT Models

Chen et al. (2021) propose SixT, a fine-tuning method for foreign-English zero-shot NMT. They initialize both the encoder and decoder to XLM-R. They follow a two-stage fine-tuning approach, first only fine-tuning the decoder layers, then continuing training by unfreezing the encoder layers and decoder embeddings. The model is trained on a parallel corpora in only de-en, and they report zero-shot to-English performance for 10 languages.

Chen et al. (2022) propose SixT+, which builds upon the authors' prior work, and is trained on a parallel corpus in 6 source languages. This is concurrent to the first submission of our work. They show their model can address zero-shot tasks from NMT to cross-lingual abstractive summarization. This work has the same goal as our first stage of training.[12] The main differences are in our training data (40 vs 6 source languages, 80M vs 120M pairs), and our simpler zero-shot training stage (no unfreezing, no position disentangled encoder).

7.2 Unsupervised MT Models

Utilizing Both Parallel and Monolingual Data Ko et al. (2021) propose NMT-Adapt, a method which follows the same data constraints as our work. Their method jointly optimizes four tasks: denoising autoencoder, adversarial training, high-resource translation, and low-resource back-translation – the latter two of which we also use. However, their work trains individual models for each direction, and furthermore for each model explicitly trains on related high-resource language datasets. This approach is thus more expensive and less adaptable to new languages as ours.

Bidirectional Multilingual NMT Garcia et al. (2021) train a single model to translate unseen languages to and from English, under the same data constraints as our work. They proceed in 3 stages, each of which uses a mixture of training data and objectives: *MASS* (Song et al., 2019) for monolingual data, *cross-entropy* for auxiliary parallel data, and both *iterative back-translation* (Hoang et al., 2018) and *cross-translation* (Garcia et al., 2020) for synthetic parallel data. This work shares our goal of developing a single bidirectional UNMT model for unseen languages. There are two main differences. First, their aforementioned training scheme is fairly involved. Second, their approach relies on cross-translation, which explicitly ties individual auxiliary languages to unseen languages, limiting their model's cross-lingual generalizability.

8 Conclusion

We have described a two-stage training approach for developing a single bidirectional, unsupervised NMT model, which we term EcXTra. The main contribution of EcXTra is in its effective synthesis of techniques from both zero-shot NMT, multilingual fine-tuning, and from unsupervised NMT, back-translation. While prior work also uses similar underlying techniques, they have much more involved training processes, either to consider the bidirectional and zero-shot direction, or introduce additional loss functions (which make training more involved). Furthermore, in this work we have taken an agnostic view towards multilinguality.

We trained a single NMT model through EcXTra, and find that each round of back-translation training further refines bidirectional translation performance. This gives rise to the view of EcXTra as successive rounds of informed initialization into further fine-tuning. The final, unsupervised EcXTra-trained model achieves competitive performance on 7 foreign-English tasks, in both directions. The straightforward nature of EcXTra allows it to be easily extended to new unseen languages.

[12] Chen et al. (2022) does perform a small-scale study on back-translation for translating English-foreign, but these models are neither multilingual nor bidirectional.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*.

Chao Bei, Hao Zong, Conghu Yuan, Qingming Liu, and Baoyong Fan. 2019. GTCOM neural machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 116–121, Florence, Italy. Association for Computational Linguistics.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 15–26, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2022. Towards Making the Most of Multilingual Pretraining for Zero-Shot Neural Machine Translation. *arXiv:2110.08547 [cs]*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Computing Surveys*, 53(5):1–38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur Parikh. 2020. A Multilingual View of Unsupervised Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3160–3170, Online. Association for Computational Linguistics.

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur Parikh. 2021. Harnessing multilinguality in unsupervised machine translation for rare languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1126–1137, Online. Association for Computational Linguistics.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Zar Zar Hlaing, Ye Kyaw Thu, Thazin Myint Oo, Mya Ei San, Sasiporn Usanavasin, Ponrudee Netisopakul, and Thepchai Supnithi. 2021. NECTEC's participation in WAT-2021. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 74–82, Online. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and Why is Unsupervised Neural Machine Translation Useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource NMT models to translate low-resource related languages without parallel data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 802–812, Online. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan

Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Anoop Kunchukuttan. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised Machine Translation Using Monolingual Corpora Only. *arXiv:1711.00043 [cs]*.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao, and Jingbo Zhu. 2019. The NiuTrans machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Shuming Ma, Jian Yang, Haoyang Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla, Alexandre Muzio, Akiko Eriguchi, Saksham Singhal, et al. 2020. Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders. *arXiv preprint arXiv:2012.15547*.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. 2022. Bidirectional language models are also few-shot learners. *arXiv preprint arXiv:2209.14500*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya. 2021. "wikily" supervised neural translation tailored to cross-lingual tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1655–1670, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, and Jie Hao. 2020. OPPO's machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 282–292, Online. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*.

Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Pétur Orri Ragnarson, Haukur Jónsson, and Vilhjalmur Thorsteinsson. 2021. Miðeind's WMT 2021 submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association* *for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Shiyue Zhang, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. 2022. How robust is neural machine translation to language imbalance in multilingual tokenizer training? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–116, Orlando, USA. Association for Machine Translation in the Americas.

A Limitations

The notable limitations are the datasets used, the compute required for training, and a want for further ablation studies.

Our training dataset *m2e-40* is a subset of the Many-English dataset (Gowda et al., 2021). This is a collection of various datasets, many of which contain mined parallel sentences. While we have assumed in our paper, like prior work, that these datasets are "real" parallel data, they are in fact quite noisy, and contain many low-quality sentence pairs that likely harm downstream system performance (Kreutzer et al., 2022).

Another potential limitation is that when we select only 2 million samples for each training language pair, instead of using all samples, we limit performance. This is possible, but our work explores a language-agnostic multilingual setting. We refer the interested reader to (Zhang et al., 2022), which finds through an empirical study that overall multilingual translation performance is best when languages are balanced.

Our method requires a solid amount of computing resources in order to train the entire NMT system (see details in Appendix C). Unlike several other works, we train a single model for all directions, which allows us to be more resource-efficient. However, very recent work has found that even without fine-tuning, multilingual pretrained LMs are able to perform zero-shot translations to and from low-resource languages (Patel et al., 2022) – so long as they are given few-shot examples (which can even be synthetic). We suspect such in-context learning based approaches will be soon popular in machine translation, as they have become in many other NLP fields.

We also note that in our work, we evaluated using only BLEU scores. BLEU, of course, is widely-used and understood in the MT community. However, over the decades, researchers have called into question relying solely on BLEU results for MT evaluation. We acknowledge this point, and keep our work as-is given our resource limitations, and given our consistency with prior unsupervised NMT work on reporting results.

A.1 Preliminary Ablations

We understand that ablation studies are useful to ascertain the contribution of various parts of the training approach. Unfortunately, we were unable to pursue this in detail because of resource limitations on our end. Therefore, we enumerate several possible ablations here, and provide preliminary observations from some small-scale experiments:

Model Size We found the large models for XLM-R and RoBERTa, instead of the base models, significantly increased performance for all language pairs and directions.

Our Dataset vs. Prior Work Datasets In the unsupervised and zero-shot NMT literature, because of the variety of task formulations and setups, works do not use consistent datasets for training. This is true for the models we provide reference comparisons to, Chen et al. (2022) and Garcia et al. (2021). These works, like ours, provide comparisons to prior work, with a disclaimer that these results cannot be completely fair. To some extent, the multilinguality agnostic dataset is a key part of the full EcXTra approach. Still, an elucidating ablation experiment could be to train our first stage model using the AUX6 dataset of Chen et al. (2022), then run back-translations using the monolingual datasets specified by Garcia et al. (2021). However, this would require additional computational resources that we unfortunately lack.

Unidirectional Unsupervised NMT We found a unidirectional English-foreign second stage model achieves similar BLEU to the bidirectional second stage models. This suggests that this MT system has no issue with bidirectionality, affirming the findings of Niu et al. (2018).

Bilingual vs. Multilingual NMT Models We found a second stage model trained to only translate a single bilingual pair, kk-en, performs quite a bit better for those translation directions than a multilingual model. This suggests that the model has difficulty with maintaining performance given all the different translation tasks, especially those with unique scripts such as Burmese and Nepali.

Training models for individual language pairs (with their own limited vocabularies), and tailoring the datasets specifically to relevant high-resource languages, is one approach as performed by (Ko et al., 2021). For example, their ne-en specific model achieves 26.3 BLEU vs. EcXTra's 8.8.[13]. However, this approach is still someone unsatisfying, as our ultimate goal is still to train a single

[13] Still, in the ne-en direction their models achieves only 18.8 BLEU (vs. EcXTra's 19.9) This suggests the multilingual similarities are currently better exploited for to-English translation, than from-English.

multilingual NMT system. We hope for continued research to close this gap between multilingual and bilingual NMT systems.

Initializing Stage 2 to Stage 1 Model In this experimental setting, we use the trained stage 1 model only to create English-foreign synthetic parallel data, but initialize to RoBERTa and XLM-R (instead of the stage 1 model). We ran this model for a few epochs, before stopping it because we found the validation BLEU increased very slowly relative to the original stage 2 training. This affirms our earlier claim that the stage 1 model is an informed initialization for the stage 2 model.

B Details on Datasets Used

Here, we expand upon Section 3 and provide further detail on the datasets used in this paper.

B.1 Zero-Shot NMT Datasets

Test We consider translation of 7 low-resource languages, which come from 6 language families. We draw these test sets from publicly available datasets from WMT21[14], FLoRes v1[15], and WAT21[16]. Where possible, we use the same test sets as specified by prior unsupervised NMT work.

Training Our first stage model is trained on a parallel dataset we term *m2e-40*. This is a subset of the Many-English[17] dataset (Gowda et al., 2021), which itself is a collection of other publicly available datasets. Of the 500 language pairs in this dataset, we choose the 40 languages with the most parallel sentences[18]. This criterion contrasts with prior work (Siddhant et al., 2022; Chen et al., 2022), which specifically select language pairs based on coverage and/or similarity to the unseen test languages. Table 5 shows more information for the training languages.

Prior work has handled the imbalance in auxiliary language pairs through temperature sampling (Devlin et al., 2019). Essentially, this is a simple trick to up-sample high-resource languages and down-sample low-resource once. In our work we take the even simpler trick of equally sampling 2 million sentences from each training language. This follows the finding of Zhang et al. (2022) that more equal sampling of languages results in the relatively best multilingual performance.

The Many-English dataset is provided as pre-tokenized and pre-processed. For our use-case, we are fine-tuning the encoder of XLM-R, which was pretrained on untokenized text. Therefore, we detokenize both the English and the foreign sides of our subset using `sacremoses`[19].

Validation The validation data comes from the development tarball of WMT19[20]. Of the 40 training languages, 15 of them are found in this tarball. As some translation directions appear multiple times (e.g. fr-en), we choose just 1 per task. Table 6 shows more information. For the supervised NMT experiment of Section 6.1, we utilize the same development datasets for the languages {zh, hi, tr, ru}.

B.2 Unsupervised NMT Datasets

Training We use several monolingual datasets for training our unsupervised NMT model. For the 7 test languages we draw from Common Crawl[21] for {kk, gu, is} and CC-100[22] for {my, ps, ne, si}.

We take the first 4M sentences of each monolingual dataset–except for Burmese (my), which has only 2M sentences. We then filter out duplicated lines, and empty lines. We thus have 26M test language sentences.

For the English-to-many direction, we require monolingual English data, which we draw from News crawl[23]. As above, we take the first 4M sentences, then filter out duplicated and empty lines. The English monolingual sentences are then translated in the 7 languages, resulting in 7 * 4M = 28M synthetic sentence pairs total.

Validation For each round of back-translation training, we use datasets in 14 directions – from/to the 7 translation directions. We withhold the first 250 sentence pairs of each translation direction (14

[14] https://www.statmt.org/wmt21/index.html

[15] https://github.com/facebookresearch/flores/tree/main/floresv1

[16] http://lotus.kuee.kyoto-u.ac.jp/WAT/my-en-data/

[17] http://rtg.isi.edu/many-eng/data-v1.html

[18] The motivation for choosing 40 languages is largely because of resource limitations on our end. Ideally, we would have liked to train on all languages with 1M+ sentence pairs.

[19] https://github.com/alvations/sacremoses

[20] http://data.statmt.org/wmt19/translation-task/dev.tgz

[21] https://data.statmt.org/ngrams/

[22] https://data.statmt.org/cc-100/

[23] https://data.statmt.org/news-crawl/en/

directions, so 3500 pairs total) to serve as validation. The early stopping criteria is standard BLEU. We tried as an alternative the round-trip BLEU proposed by Lample et al. (2018), but found this made little difference in final evaluation results.

C Modeling and Training Setup

Our research was pursued in a resource-limited setting. For training, we used 4 NVIDIA RTX A6000 GPUs (48GB vRAM each). For inference, we used the above, and additionally had access to 16 NVIDIA GeForce RTX 2080 Ti GPUs (11GB vRAM each).

Given the above resource-limited training and inference setup, we provide some rough estimates of runtime. Training a stage 1 model takes about 1 week. Training a stage 2 model takes about 6 weeks, given the steps: a) run xx->en back-translations on 26m sentences (2 weeks), b) train the round 1 model (1 week), c) run en->xx back-translations on 28M sentences (2 week), d) train the round 2 model (1 week). Given more standard GPU resources, we would expect at least a 3-4x speedup in the whole training process.

We use the `transformers` package (Wolf et al., 2020) as the backbone for our modeling work. Specifically, we use it to load pretrained model weights and tokenizers. The rest of the code is implemented in PyTorch (Paszke et al., 2019).

Hyperparameters The most up-to-date version of the hyperparameters can be found in the repository.[24] For training, the batch size = 20000 for round 0, and 11500 for rounds 1 and 2. We use an Adam optimizer, with learning rate = 1e-3, and warmup steps = 12500. The learning rate decay schedule is based on the inverse square root of the update number. The dropout probability = 0.1, and the random mask probability = 0.4. For inference, the batch size = 1500, and beam size = 5.

D Start Tokens to Indicate Target Language

Following Johnson et al. (2017), we add special start tokens to each source sentence, to indicate the desired target language. This only applies to stage 2, because stage 1 always targets English. The default implementation directly adds these tokens, of the form `<2xx>` to the target vocabulary. Our setting requires adapting the implementation because as we have frozen the target embeddings (and source embeddings), we cannot increase the vocabulary size. We therefore indicate the target language with a two-token sequence, which consists of the usual start token `<s>`, and another token $\texttt{TOK}_\texttt{i}$ drawn from the long tail of the vocabulary. The model then must learn that `<s> +` $\texttt{TOK}_\texttt{i}$ means to translate to a given language.

To be concrete, we use XLM-R tokenization, which consists of 250,002 SentencePiece tokens. For this paper, in which the model supports 8 languages, we arbitrary select indices 202201 to 202208, and assign each to a language.

E How Zero-Resource is Zero-Resource?

In this work, we have defined zero-resource as the setting in which no parallel sentences are available for a language pair of interest. This definition follows the general usage in the field. To be exactly precise, though, the pretrained multilingual model used, XLM-RoBERTa, has indeed seen monolingual text in each of the 7 low-resource languages.

F Sample Output

Sample output for the EcXTra NMT models are shown in Tables 7 and 8.

[24] https://github.com/manestay/EcXTra/

Code	Language	Family	Script	Source	# Pairs
kk	Kazakh	Turkic	Cyrillic	newstest2019	1000
gu	Gujarati	Indic	Gujarati	newstest2019	1016
si	Sinhala	Indic	Sinhala	FLoRes v1	2905
ne	Nepali	Indic	Devanagari	FLoRes v1	2924
ps	Pashto	Iranian	Arabic	newstest2020	2719
is	Icelandic	Germanic	Latin	newstest21	1000
my	Burmese	Burmese-Lolo	Burmese	WAT21	1018

Table 4: Information for the **test** languages, and the foreign-English datasets used. The columns are, from left to right, the ISO 639-1 language code, the name of the language, the language family at the Genus level, the data source, and the number of sentence pairs.

Code	Language	Code	Language
tr	Turkish	hu	Hungarian
sr	Serbian	sl	Slovenian
fr	French	vi	Vietnamese
he	Hebrew	et	Estonian
ru	Russian	sk	Slovak
ar	Arabic	ja	Japanese
zh	Chinese	lt	Lithuanian
bs	Bosnian	lv	Latvian
nl	Dutch	uk	Ukrainian
de	German	th	Thai
pt	Portuguese	cs	Czech
no	Norwegian	ko	Korean
it	Italian	id	Indonesian
es	Spanish	ca	Catalan
pl	Polish	mt	Maltese
fi	Finnish	ro	Romanian
fa	Persian	bg	Bulgarian
sv	Swedish	hr	Croatian
da	Danish	hi	Hindi
el	Greek	eu	Basque

Table 5: Information for the **train** languages. The columns are, from left to right, the ISO 639-1 language code, and the name of the language.

Code	Language	Source	# Pairs
tr	Turkish	newsdev2016	1001
fr	French	newstest2009	2525
ru	Russian	newstest2012	3003
zh	Chinese	newsdev2017	2002
de	German	newstest2009	2525
it	Italian	newstest2009	2525
es	Spanish	newstest2009	2525
fi	Finnish	newsdev2015	1500
hu	Hungarian	newstest2009	2525
et	Estonian	newsdev2017	2000
lt	Lithuanian	newsdev2019	2000
lv	Latvian	newsdev 2017	2003
cs	Czech	newstest2009	2525
ro	Romanian	newsdev2016	1999
hi	Hindi	newsdev2014	520

Table 6: Information on the **validation** languages, and the foreign-English datasets used. The columns are, from left to right, the ISO 639-1 language code, the name of the language, the source (from WMT development set), and the number of sentence pairs.

Model	Translation (kk-en)
Reference	The first medal place was given to Dastan Aitbay from Kyzylorda and his project on "Safe Headphones" Innovative headphones".
EcXTra-r_0	The winning first place was won by Dastan Aitbay's innovative earpiece "Safe headphones" from the city of Kyushu.
EcXTra-r_1	First place was won by Dastan Attbay of the city of Kyrgyzlord "Innovative earphones "Safe headphones."
EcXTra-r_2	The cool first place was won by Dastan Aitbay, from the city of Kyrgyzstan, the "Inventive earcap Safe Headphones."

Table 7: Sample kk-en unsupervised translations for the input: Жүлделі бірінші орынды Қызылорда қаласынан Дастан Айтбайдың "Инновациялық құлаққап "Safe headphones"жобасы жеңіп алды.

Model	Translation (en-is)
Reference	Markmiðið er að fegra svæðið og leyfa mósaíkverki Gerðar Helgadóttur á Tollhúsinu að njóta sýn betur.
EcXTra-r_0	N/A
EcXTra-r_1	Markmið er að fagna svæðið og gera mosaík Gerður Helgadóttir á Tollhúsinu áberandi.
EcXTra-r_2	Tilgangurinn er að fallega svæðið og gera mosamynd Gerður Helgadóttir á tollhúsinu meira áberandi.

Table 8: Sample en-is unsupervised translations for the input: The aim is to beautify the area and make Gerður Helgadóttir's mosaic on the Customs House more prominent.