Decoupled Vocabulary Learning Enables Zero-Shot Translation from Unseen Languages

Carlos Mullov[1] and **Ngoc-Quan Pham[1]** and **Alexander Waibel[1,2]**
[1]Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Carnegie Mellon University, Pittsburgh PA, USA
`{firstname}.{lastname}@kit.edu`

Abstract

Multilingual neural machine translation systems learn to map sentences of different languages into a common representation space. Intuitively, with a growing number of seen languages the encoder sentence representation grows more flexible and easily adaptable to new languages. In this work, we test this hypothesis by zero-shot translating from unseen languages. To deal with unknown vocabularies from unknown languages we propose a setup where we decouple learning of vocabulary and syntax, i.e. for each language we learn word representations in a separate step (using cross-lingual word embeddings), and then train to translate while keeping those word representations frozen. We demonstrate that this setup enables zero-shot translation from entirely unseen languages. Zero-shot translating with a model trained on Germanic and Romance languages we achieve scores of 42.6 BLEU for Portuguese-English and 20.7 BLEU for Russian-English on TED domain. We explore how this zero-shot translation capability develops with varying number of languages seen by the encoder. Lastly, we explore the effectiveness of our decoupled learning strategy for unsupervised machine translation. By exploiting our model's zero-shot translation capability for iterative back-translation we attain near parity with a supervised setting.

1 Introduction

In order to extend Machine Translation from a hundred languages to the 7,000 languages of the world new methods have to be developed with more economical use of data from individual languages. This requires better use of abstractions and representations of languages to facilitate transfer learning to low-resource languages. In machine translation, it had been attempted by developing linguistic structures explicitly in classical Interlingua systems (Levin et al., 1998) or imposing structure on statistical models (Suhm, 1994; Wang and Waibel, 1998).

However, developing such structures was labor intensive and did not achieve acceptable performance or required domain restrictions. First attempts toward learning latent representations of language (Wang and Waibel, 1991, 1995) still required linguistic structures or specific design. Successful performance on open-domain translation was only feasible later using multilingually trained end-to-end neural translation models (Bahdanau et al., 2015; Johnson et al., 2017; Ha et al., 2016). These models share hidden representations between languages and allow for multilingual translation that benefits low-resource languages by transfer learning from high-resource languages and enable zero-shot translation on new language directions.

The multilingual encoder in such models will ideally learn to map semantically similar sentences onto similar latent representations – even across different languages. Ongoing research in the field of zero-shot translation shows that multilingual NMT models exhibit this property to a certain extent (Duquenne et al., 2023) and that enforcing the similarity between sentence representations across different languages in turn also improves the zero-shot capabilities (Pham et al., 2019; Arivazhagan et al., 2019; Zhang et al., 2020; Liu et al., 2021; Pan et al., 2021). At the same time, cross-lingual transfer learning research shows that massively multilingual systems can be rapidly extended to new languages on very little data (Neubig and Hu, 2018; Artetxe et al., 2020; Garcia et al., 2021; Marchisio et al., 2023; Chen et al., 2023). This suggests that exposing the NMT model to a higher number of languages increases the plasticity of its sentence representation and enables easier adaptation to unseen languages.

When extending a multilingual model by an unseen language, however, naturally the question arises of how to deal with the unknown vocabulary words. The field of incremental learning devised several strategies for adapting to new vocabularies.

Artetxe et al. (2020) show that simply relearning the embedding layer can be done efficiently, compared to training the model from scratch. Garcia et al. (2021) find that retraining the shared BPE codes once the new language data become available allows one to keep most of the pre-trained vocabulary, making even more efficient adaptation possible.

In this work, we go one step further and show that by designing a system where word representations are learned in a separate step we can zero-shot translate from an unknown language without any adaptation whatsoever. We therefore employ cross-lingual word embeddings (Artetxe et al., 2016; Lample et al., 2018b) – a technique that has extensively been used in unsupervised machine translation, but hasn't so far been given much consideration for incremental learning – to align monolingual word representations into a common space, and then use these as our regular word embedding layer in the NMT system. Crucially, keeping these embeddings frozen in the NMT training allows us to align new language vocabularies into the common embedding space post-training.

Furthermore, we show that this zero-shot translation capability allows us to easily generate synthetic parallel data for unknown languages and enables simple and efficient unsupervised machine translation without the need for denoising autoencoding on massive amounts of monolingual data.

Finally, we also experiment on continuous output NMT (Kumar and Tsvetkov, 2019), which – while underperforming the softmax models in terms of BLEU scores – perfectly fits our setup and attains vastly superior performance in terms of training times.

We make our code and training configurations available under https://github.com/cmullovisl/clwe-transfer and we release our pre-trained multilingual models and embedding alignments.

2   Multilingual Neural Machine Translation

In training the multilingual NMT system we aim to estimate the probability $\mathbb{P}(Y_{\ell_{tgt}} = y \mid X_{\ell_{src}} = x)$ that the sentence y in the target language ℓ_{tgt} is a suitable translation of the source sentence x in the source language ℓ_{src}. We describe the distribution of the sentences in different natural languages ℓ through random variables X_ℓ, Y_ℓ. A universal encoder maps the sentences from the different input distributions onto a single, shared latent distribution $H_{\text{enc}} = enc(X_\ell)$. A decoder is then tasked to model the probability distribution from this latent variable: $\mathbb{P}(Y_{\ell_{tgt}} = y \mid H_{\text{enc}} = c)$. Identically to a regular bilingual NMT system, the multilingual translation system consists of a neural network, which, in a supervised setting, is trained to maximize the likelihood of a dataset of parallel sentences $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where sentence pairs (x_i, y_i) are translations of each other:

$$\min_{\vartheta \in \Theta} \left\{ - \sum_{i=1}^{n} \log \mathbb{P}\left(X_{\ell_{tgt}^i} = y_i \mid X_{\ell_{src}^i} = x_i, \vartheta \right) \right\}$$

Note that – aside from the added target language selection mechanism – the only difference to bilingual translation is the nature of the input and output distribution. Thus, the main practical difference in the training a multilingual system is that instead of using a bilingual training corpus, $\mathcal{D}$ consists of a concatenation of several parallel corpora.

3   Related Work

Cross-Lingual Word Embeddings   Qi et al. (2018) look into the effectiveness of pre-trained embeddings for regular supervised NMT and find them to be helpful in a multilingual setting. Kim et al. (2019) conduct unsupervised MT by transferring the parameters learned on a high resource language pair to a new source language. Similar to our work they therefore align the new language embeddings to the parent model source embeddings, but they focus on a bilingual setting. Compared to their work we show that this setup not only allows for unsupervised MT but zero-shot translation from unseen languages, especially when combined with a multilingual setup.

Unsupervised MT   Our work is closely related to unsupervised NMT (Lample et al., 2018a,c; Artetxe et al., 2017, 2019; Conneau et al., 2020; Song et al., 2019; Liu et al., 2020b). Initial works train unsupervised NMT systems by bootstrapping a common representation space between two languages through a combination of cross-lingual word embeddings and denoising autoencoding, and then exploit these common representations for iterative back-translation (Lample et al., 2018a). Through unsupervised means they thus induce a shared sentence representation, essentially bootstrapping a multilingual NMT system from monolingual data.

Later works do away with the cross-lingual word embeddings and create this shared representation space through large scale denoising autoencoding on massive amounts of monolingual data (Song et al., 2019; Liu et al., 2020b) and multilinguality (Sen et al., 2019; Siddhant et al., 2020; Conneau et al., 2020; Liu et al., 2020b). In contrast to this development, we show that cross-lingual word embeddings remain valuable when combined with regular NMT training, allowing to bootstrap high-quality multilingual representations even on small-scale datasets.

New Language Learning Multiple works look into adding unseen languages to multilingual encoder-decoder systems. Neubig and Hu (2018) concerned about the possibility of rapidly extending a multilingual NMT model by an unseen low-resource language. They compare between bilingual training, multilingual training alongside a similar source language and multilingual pre-training with as many languages as possible. They come to the conclusion that pre-training in a highly multilingual setting – in their case 58 source languages – significantly improves transferability to low-resourced languages. In particular, they achieve significant performance (up to 15.5 BLEU) on an entirely unseen language, provided parallel training data on a closely related language. In this work we use the same parallel data for NMT training and significantly improve upon their result through the employment of cross-lingual word embeddings. Escolano et al. (2019) devise an approach to multilingual NMT with independent encoders and decoders that enables addition of new languages. Li et al. (2023) fine-tune a model based on XLM-R and use it to zero-shot translate from languages seen in pre-training, but unseen in the supervised downstream task. Tang et al. (2020) look into extending mBART25 (Liu et al., 2020b) by 25 new languages and show that given enough modelling capacity one can add new languages without loss in quality on existing ones. In their work, however, they only consider new languages sufficiently represented in the base model vocabulary, and they leave the original vocabulary unmodified.

Vocabulary Adaptation Kim et al. (2019), Garcia et al. (2021), Marchisio et al. (2023) and Chen et al. (2023) look into vocabulary adaptation for new languages. Garcia et al. (2021) show that rebuilding the subword vocabulary once the new language data is available results in most of the vocabulary matching the old one. Retaining the embedding vectors for common subword units allows for quick adaptation and minimal forgetting of the existing languages. In our work we directly compare to this method and show that a vocabulary based on aligned pre-trained embeddings not only presents a viable alternative to a subword-based one, but also allows for more efficient adaptation to and zero-shot translation from unseen languages.

4   Generalizing to Unseen Languages

To learn a multilingual NMT system that easily generalizes to unseen languages we propose a decoupled learning approach, divided into multiple steps:

1. train word embeddings on monolingual data for each of the *base languages* $\ell \in \mathcal{L}_{\text{base}}$

2. align the word embeddings into a common embedding space

3. initialize the encoder embedding layer of a standard NMT model with the learned word representations

4. supervised NMT training on $\mathcal{L}_{\text{base}}$ parallel data with frozen encoder word embeddings

In this setup we may later extend the translation model by an unseen language $\ell_{\text{new}} \notin \mathcal{L}_{\text{base}}$ by learning word representations for ℓ_{new} and aligning them into the common word embedding space (see Section 4.1). We show that this alignment alone enables us to start translating from ℓ_{new} in a plug-and-play fashion, even without any additional updates to the NMT model's main body (i.e. the Transformer layers).

Optionally, the learned word representations may also be employed on the decoder-side, which will further enable the back-translation-based adaptation detailed in Section 5.3. Figure 1 describes our decoupled learning NMT architecture.

4.1   Decoupled Vocabulary Learning

In conventional training of a universal multilingual NMT system the ℓ_{new} word vectors are randomly initialized and then trained end-to-end on the NMT objective. In training the model then learns to represent its words in a shared multilingual embedding space, by learning cross-lingual word correlations from the parallel data. In a monolingual data only setting we aim to emulate these multilingual word representations through cross-lingual word embeddings. These word embeddings are then integrated
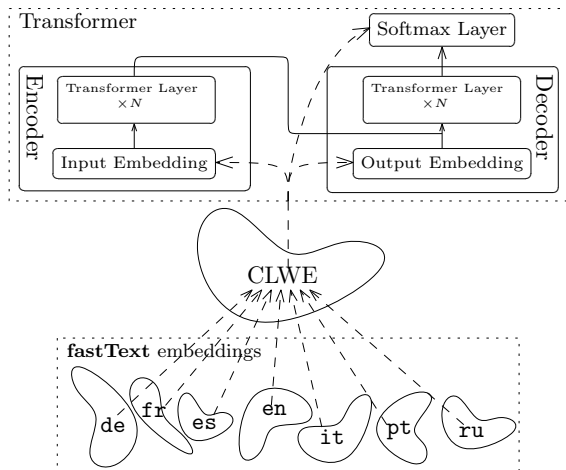
Figure 1: Our NMT architecture consists of a Transformer model with pre-trained cross-lingual word embeddings (CLWE) for embedding layers. The embedding vectors are obtained through alignment of monolingual *fastText* embeddings for each language into a common embedding space.
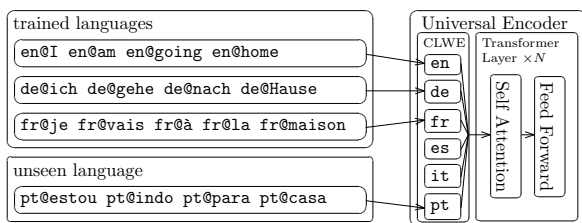

Figure 2: Our decoupled learning of word representations enables us to zero-shot translate from an unseen language ℓ_{new} in a plug-and-play fashion. The Transformer layers have no prior exposure to ℓ_{new}. The cross-lingual word embeddings (CLWE) serve as lookup table for ℓ_{new} words.

into our encoder-decoder NMT model by simply using them in the word embedding layers.

For each language $\ell \in \mathcal{L}_{\text{base}} \cup \{\ell_{\text{new}}\}$ we use pre-trained monolingual word embeddings learned with *fastText* (Bojanowski et al., 2017), which we then align into one common space through cross-lingual embedding alignment (Joulin et al., 2018). For a common alignment between multiple languages we choose to align into a *single hub space* (SHS Heyman et al., 2019), i.e. we pick one of our base languages as a *pivot* and align each of the embedding spaces E_ℓ to E_{pivot}.

4.2 Zero-Shot Translation from Unseen Languages

To provide an indication of how well the universal encoder generalizes an unseen language ℓ_{new} we let our multilingual model decode from ℓ_{new}

sentences, without any sorts of prior exposure to ℓ_{new}. As outlined in Figure 2 we therefore employ the cross-lingual embeddings to look up the ℓ_{new} word vectors and then simply decode from the resulting sequence. While the encoder may not know the syntax of the unseen language, many syntactical concepts in language – such as grammatical case, part-of-speech, grammatical genders, or tense – are encoded at the word level[1]. As such we hypothesize that the word vectors trained on the new language alone provide enough syntactic level information to perform the language comprehension task to a certain degree.

To test our hypothesis that a high degree of multilinguality will help in making the sentence representations more general we can then vary the number of languages the encoder gets exposed to in training. We may then observe how this affects languages closely related to languages in $\mathcal{L}_{\text{base}}$ as well as languages that are distant to any of the languages in $\mathcal{L}_{\text{base}}$.

4.3 Unsupervised Machine Translation

Iterative back-translation (Sennrich et al., 2016; Lample et al., 2018a; Artetxe et al., 2018) is a common approach in unsupervised NMT, where a bidirectional model is used to generate synthetic parallel data and then subsequently adapted on this data to then generate even better synthetic data. To iteratively improve translation between two languages ℓ_1 and ℓ_2 one would

1. translate ℓ_1 monolingual data to ℓ_2 to generate $\ell_2 \to \ell_1$ parallel data
2. adapt the model on the $\ell_2 \to \ell_1$ data
3. translate ℓ_2 monolingual data to ℓ_1 to generate $\ell_1 \to \ell_2$ parallel data
4. adapt the model on the $\ell_1 \to \ell_2$ data

The resulting model will be able to produce better $\ell_2 \to \ell_1$ synthetic data than in Step 1 and the process may be repeated for iterative improvements until convergence.

To obtain usable translations for Step 1 the model, however, must first be bootstrapped, which is usually done through denoising autoencoding or language modelling on massive amounts of monolingual data. This process is both costly and oftentimes fails for languages that are either too distant

---

[1]Take German for example, a noun would have different surface forms for different grammatical cases, which the fastText model should take into account and appropriately assign different word vectors.

or have too little monolingual data available (Kim et al., 2020; Marchisio et al., 2020). We show that in our setup we may skip this step entirely by zero-shot translating from an unknown language ℓ_{new} to kick-start the iterative back-translation process. Furthermore, we show that through our model's improved ability to generalize to unseen languages we can robustly improve ℓ_{new} translation quality, even when ℓ_{new} is so distant to any other language $\ell \in \mathcal{L}_{\text{base}}$ that the zero-shot generated synthetic data is barely intelligible, and even on relatively small amounts of monolingual ℓ_{new} data.

5 Experiments

5.1 General Experiment Setup

Cross-lingual word embeddings As the basis for all of our monolingual word embeddings we use pre-trained *fastText* models[2] trained on Common Crawl (Wenzek et al., 2020). These models provide 300-dimensional high-quality embeddings for full word units, with pre-trained models available in 157 languages. Using the pivot approach described in section 4.1 we align these embeddings into a common embedding space using direct supervised optimization on the RCSLS criterion (Joulin et al., 2018) using the bilingual dictionaries provided by the MUSE project (Lample et al., 2018b). To reduce the vocabulary size, we subsequently regenerate our vocabulary and the corresponding embedding vectors using only the words in our NMT training corpus. For alignment accuracies and vocabulary sizes see Appendix Table 3.

Full word-units While using a joint BPE lets us share information between vocabularies and reduces the size of the multilingual vocabulary, for our approach we opt for full word units. Our preliminary experiments have shown full-word units to outperform our BPE models when used in combination with frozen cross-lingual embeddings. Furthermore, this gives us better compatibility with distant languages with separate alphabets and avoids the problem of having to extend the learned BPE codes post-training when we add a new language to the model. To alleviate the OOV word problem we use *fastText* as the basis for our monolingual word embeddings, which, by using subword level information, gives us the ability to map unseen words into the embedding space.

From an implementation perspective we merge our monolingual vocabularies and concatenate their embedding matrices, while dealing with duplicate words across different languages by encoding each word with a language specific prefix, e.g. we encode the English word *bank* as **en@bank**. At inference-time we limit the target vocabulary to the language we decode to, thereby forcing the model to only choose tokens from our desired target language, which is not trivial in a model based on BPE units shared across multiple languages. Due to the large vocabulary sizes using full word-units increases the number of parameters immensely. However, since we keep the embedding layers frozen the amount of trainable parameters and GPU memory consumption[3] *decreases* in practice. Furthermore, our approach results in lower sequence lengths, thus partially making up for higher computational complexity of the larger softmax embedding layer. At inference-time (and potentially also training time) the computational complexity may further be reduced through dynamic vocabulary sub-selection (Senellart et al., 2018). We leave this line of research for future work.

Von Mises-Fisher loss Since the fastText embeddings we employ in the output vocabulary are normalized to unit norm, the matrix multiplication in the final projection layer is similar to a cosine-similarity search through the fastText embedding space. Due to our large vocabulary sizes this makes the computation of the output layer during training and test time more costly. To counter this issue we conduct experiments with von Mises-Fisher loss for NMT (Kumar and Tsvetkov, 2019), where model outputs are optimized toward frozen pre-trained word embedding targets, which perfectly fits our setup. This loss is similar to the cosine-loss and it makes the computational complexity of training entirely independent of the vocabulary size, and we observe up to $5.8\times$ speedups in training throughput. We find this approach to consistently under-perform the traditional softmax-based models (Appendix C), but in light of the speedup we keep this approach for the Section 5.2 experiments. In the following we refer to this model variant as vMF.

Datasets As our training, development and evaluation data sets we use multilingual transcriptions

---

of TED talks in 60 languages (Qi et al., 2018). For our main body of experiments we include English, German, Spanish, French and Italian as the supervised languages, which we refer to as *Romance-Germanic* languages. We train the basesystem on 20 translation directions, and a parallel corpus size of $3,576,046$ sentences. For our experiments on unsupervised MT we focus on monolingual in-domain data obtained from the same corpus of TED talks. To prevent contamination this corpus is split into train, validation and test data using the same TED talks across all languages (Qi et al., 2018). For larger scale out-of-domain experiments we use monolingual data from the Wiki-40B dataset (Guo et al., 2020) and newscrawl data (Bojar et al., 2014) and we further evaluate on Flores (Goyal et al., 2022) and Tatoeba (Tiedemann, 2020).

NMT model For our NMT model we use a Transformer model (Vaswani et al., 2017) with relative position encodings (Shaw et al., 2018) and post-layernorm. In accordance with our 300-dimensional fastText embeddings we train 9-layer models with embedding dimension 300 unless otherwise specified. Our implementation is based on the repository provided by Kumar and Tsvetkov (2019)[4], which is based on *OpenNMT-py* (Klein et al., 2017). We share embedding parameters across the encoder and the decoder wherever applicable, and use tied output embeddings. Refer to Appendix B for the full set of training parameters.

We train our softmax-based models using Adam (Kingma and Ba, 2015) and our vMF model (Kumar and Tsvetkov, 2019) using RAdam (Liu et al., 2020a) for 160,000 updates or until early stopping. For estimates of training times and GPU hours see Appendix G. We evaluate our models using the ChrF++ metric (Popović, 2017) and list BLEU (Papineni et al., 2002) scores in the appendix. The supervised direction evaluation scores for our Romance-Germanic models are listed in Table 5.

Finally, to verify the viability of our setup – namely a full word-unit translation system with frozen pre-trained word embeddings – we also train a standard subword-unit translation system which matches our proposed model as closely as possible. We train the BPE model on the concatenated multilingual corpus for 50k merge operations. The results show an average score of 55.1/51.5/55.3 ChrF++ for the softmax-based model, the vMF

---

model and the BPE baseline (detailed in Table 5), demonstrating that while the vMF model lags behind in terms of translation performance, the softmax-based model closely matches the baseline despite the word embeddings not being optimized for translation.

To ensure better reproducibility we base our data and training pipelines on DVC (Kuprieiev et al., 2024). To avoid data leakage of unseen languages into supervised training we structure our experiments in two stages, where the choice of ℓ_{new} may be postponed until the second stage.

5.2 Zero-Shot

In the following we present our results for translating from unseen languages after aligning their word embeddings into the common word embedding space and plugging them into the NMT model's encoder without any further adaptation. We report the results in Table 1. For comparison, we also evaluate the translation performance of a regular BPE-based model on unseen languages and observe that it fails to produce reasonable translations. Similar to the findings of Neubig and Hu (2018) we observe up to 12.9 BLEU (34.9 ChrF++) on Portuguese-English, but find the translations to be unintelligible.

For additional comparison we employ Garcia et al.'s (2021) method of adapting the model via vocabulary substitution. Matching their findings we observe that rebuilding the subword vocabulary with the new language results in 77 - 95 % overlap with the old vocabulary and after adaptation the translation performance nearly matches the oracle setup. We further improve on their method by freezing the decoder (but training the cross-attention weights) when adding a new source language, improving by +0.8 ChrF++ (+0.6 BLEU) on Portuguese-English. Compared to the resulting 62.4 ChrF++ after adaptation we achieve up to 61.0 ChrF++ translating zero-shot on our in-domain test set. On the out-of-domain test sets we even outperform the adapted system by +3.1 ChrF++ on Flores (27.5 vs 30.6) and +1.8 ChrF++ on Tatoeba (36.9 vs 38.7). Note, that this best performing model is our Romance-Germanic model with several supervised languages close to Portuguese. For the same model we observe much higher gaps between supervised and zero-shot performance: For Russian the baseline out-performs our method by 4 ChrF++, which is, however, reasonable considering that our model has at the time of decoding never been exposed to any Slavic language.

Trained on	$\|\mathcal{L}_{\text{base}}\|$	d_{model}	de-en	es-en	ru-en	vi-en	pt-en	nl-en	bg-en	ar-en	tr-en	ko-en
Softmax Models												
BPE baseline	5	300	57.9	62.7	11.0	12.4	34.9	23.4	11.4	9.9	16.1	11.1
(Garcia et al., 2021)	5	300			47.4	48.9	62.4	57.3	59.2	50.9	48.2	41.7
en,de,es,fr,it	5	300	57.5	62.6	42.4	35.1	61.2	51.8	49.9	38.8	32.5	25.1
en,de,es,fr,it,ru,uk,pl,vi,zh	10	576	56.1	61.5	48.1	48.2	60.4	51.2	53.2	41.7	34.8	25.0
supervised fine-tune	5	300			48.5	46.9	63.3	56.5	58.5	52.6	48.2	39.1
vMF Models (Kumar and Tsvetkov, 2019)												
es→en	1	300	36.0	60.9	35.1	27.5	54.7	40.5	33.9	31.2	26.9	21.6
es↔en	2	300	40.9	59.9	37.1	30.0	55.2	42.4	42.1	34.3	29.0	22.6
en,de,fr,it	4	300	54.3	55.1	39.6	30.1	56.0	47.6	45.5	33.8	29.2	21.4
en,de,es,fr,it	5	300	54.3	59.5	40.0	30.8	58.1	47.9	45.7	34.9	29.4	20.4
en,de,es,fr,it,ru,uk,pl	8	300	52.7	57.8	45.2	32.8	56.4	47.9	50.5	37.1	31.2	22.3
en,de,es,fr,it,ru,uk,pl,vi,zh	10	300	51.9	57.2	44.6	45.4	55.2	47.5	49.4	37.0	31.0	23.0
supervised fine-tune	5	300	54.4	59.9	45.5	44.3	60.4	53.5	56.3	48.4	47.9	33.4

Table 1: ChrF++ scores for decoding from an unseen language on the TED data test split (Qi et al., 2018). Grey colored cells indicate supervised training on the source language, white cells indicate no prior exposure to the source language. d_{model} denotes the Transformer model dimension and arrows indicate what models are used as basis for fine-tuning. Note, that every column in the three fully supervised rows represents a separate bilingual fine-tune. For BLEU scores and translation directions other than English see Appendix D.

With growing distance to the supervised languages this gap grows higher. For Korean the English translations end up being barely intelligible[5]. Nevertheless, the translations are still good enough to kick-start the iterative back-translation process, as we show in Section 5.3. We observe that on more distant languages the model oftentimes gets stuck in decoding loops, which we address in a post-processing step to remove repeated n-grams.

Effect of Multilinguality To study the effect of multilinguality we train multilingual models on varying numbers of languages and families and compare the translation performance on unseen languages in Table 1. On the supervised languages we see a clear trend of performance degradation the more languages we include in our model, e.g. for Spanish→English we see clear degradation when going from bilingual to multilingual. For the zero-shot directions on the other hand we see a clear boost in performance when going from bilingual to multilingual. As we increase the number of languages, however, we do not see a clear trend of im-

provement. Rather than the number of languages, the relatedness of the added languages plays the more important role, i.e. adding non-Romance to the model continuously degrades performance on Portuguese, and on Slavic languages the performance increases once we include Slavic languages in training and decreases once we include Asiatic languages.

Finally, for the three languages aren't directly related to any of the supervised languages (i.e. Arabic, Turkish, Korean) we see a jump in performance upon inclusion of Slavic languages, but no consistent improvement on inclusion of Asiatic languages.

5.3 Unsupervised Machine Translation

In Section 5.2 we show that our setup enables translation from unknown languages, albeit with lower quality the further the distance from our training languages. In our final experiment we exploit this capability for zero-shot iterative back-translation (Lample et al., 2018a). We therefore simply perform beam search to translate the monolingual target language portion of our TED corpus to each of the known languages to generate synthetic parallel data. On this synthetic data we then train

---

[5]Note, that we have observed some tokenization issues with Korean, leading to large vocabulary size of 330k. Generally, we observe drops in performance when fastText requires special tokenization, i.e. for Chinese, Japanese, Vietnamese.

Model	en→pt			en→ru			en→tr			en→ko		
	TED	Flores	Tatoeba	TED	Flores	Tatoeba	TED	Flores	Tatoeba	TED	Flores	Tatoeba
(Song et al., 2019)							23.4	29.7	20.0			
(Artetxe et al., 2019)							10.8	12.1	12.5			
Iteration 1	59.6	53.3	56.8	41.4	39.9	48.4	35.1	34.7	35.9	15.1	12.8	13.6
Iteration 3	60.0	53.7	57.2	42.9	41.9	50.9	38.7	37.7	42.4	18.3	14.7	15.6
Iteration 5	60.0	53.8	57.2	43.3	42.5	52.9	40.4	38.9	45.6	19.5	15.3	16.3
Supervised	59.0	52.8	56.0	42.8	41.5	51.3	42.2	39.8	48.6	20.6	15.2	17.4

Model	pt→en			ru→en			tr→en			ko→en		
	TED	Flores	Tatoeba	TED	Flores	Tatoeba	TED	Flores	Tatoeba	TED	Flores	Tatoeba
(Song et al., 2019)							24.5	31.0	20.5			
(Artetxe et al., 2019)							8.7	9.3	8.1			
Zero-shot	61.2	56.3	60.5	42.4	40.9	46.0	32.5	32.9	28.4	25.1	25.9	21.3
Iteration 2	62.6	57.4	62.4	46.1	45.5	52.1	40.4	40.0	41.0	31.8	31.8	27.4
Iteration 4	62.4	57.5	62.3	47.1	47.3	55.4	43.0	41.7	46.7	33.1	33.3	29.5
Iteration 6	62.3	57.6	62.1	47.1	47.4	55.8	44.5	42.7	49.9	33.6	33.8	30.5
Supervised	63.3	56.7	62.3	48.5	46.5	57.3	48.2	44.2	54.4	39.1	34.8	40.1
# of Sentences	52k			208k			182k			206k		

Table 2: ChrF++ scores after each iteration of back-translation. In each iteration we use our Romance-Germanic multilingual model to zero-shot translate from TED domain to create synthetic parallel data.

our softmax-based Romance-Germanic model for 16,000 steps or until saturation. In training, we freeze the encoder in each odd iteration, the decoder in each even iteration. Table 2 reports the ChrF++ scores after each iteration of training on back-translated data (Table 9 for BLEU scores), alongside the supervised fine-tune on real bilingual data. We observe that for the closest language (Portuguese) a single round of back-translation closes the gap, and on the out-of-domain test set even outperforms the supervised fine-tune. On the more distant languages we see a rapid closing of gap between the supervised and the unsupervised fine-tune.

Note that our primary goal is to demonstrate that our setup enables us to skip the expensive bootstrapping process involving denoising autoencoding on massive amounts of monolingual data exhibited in most approaches to unsupervised MT. As such we don't continue the iterative back-translation process until saturation, but execute at most 6 iteration. We believe that more iterations will further close the gap between supervised and unsupervised performance.

Effect of Domain Many analyses of unsupervised MT have shown the importance of the considered domain. As the Table 2 bottom row shows the monolingual data we use for our in-domain experiments ranges from 52k to 208k sentences, which can be considered extremely small amounts when it comes to monolingual data. To study the effect of domain and amount of data we perform training on Wikipedia and news domain in Appendix F. We observe that changing the domain helps in some cases, for improving the overall translation performance a larger number of iterations on TED domain, however, seem to be most effective.

Finally, we compare to the approaches detailed in Song et al. (MASS 2019) and Artetxe et al. (2019). For both systems we use their provided implementations to train an English-Turkish unsupervised MT system, using 1947k sentences of Turkish Wikipedia data for bootstrapping. Both systems end up failing to produce any reasonable translations on any of the test sets. To verify our setup we further use MASS to train an English-French unsupervised MT model on the French portion of our Wikipedia data. This model ends up achieving 52.4 (54.1) ChrF++ from (to) English on the Flores test data.

6 Conclusion

In this work, we have looked into the generalization ability of a multilingual NMT system on unseen languages. We therefore propose a decoupled learning setup for MT training, where we learn the word representations of each language in a separate step. This enables easier generalization to new languages as this approach enables extension of the vocabulary post-training. We show that through simple alignment of an unseen language word representations into the common vocabulary space, we can produce intelligible translations without any further adaptation. We further show that on languages similar to the ones the NMT system has already been exposed to, this zero-shot translation setup even approaches supervised translation performances.

Lastly we show that this capability to produce translations for unseen languages can be exploited for efficient unsupervised machine translation.

Future Work In the future we would like to explore the efficacy of our presented setup for an incremental learning system, which would entail in-depth experiments on catastrophic forgetting.

7 Limitations

The method described in this work relies on high quality word embeddings and alignments between them. Training such high-quality word embeddings requires access to large amounts of monolingual data, which may be available in abundance for high and mid-resource languages, but may be unavailable in the amounts necessary for endangered languages, or languages which are primarily spoken, such as dialects or languages without an official orthography. This issue might be dealt with through data-efficient means of training word embeddings, which might entail an adaptation of word embeddings from a related high-resource language. Furthermore, we rely on bilingual dictionaries for learning high-quality alignments between embedding spaces. This may present a limitation for low-resource scenarios, in practice, however, bilingual dictionaries are less of an issue than the availability of monolingual data (see Kamholz et al., 2014; Vulić et al., 2019).

In part due to the availability of high-quality multi-parallel data and in part due to a limit in computational resources our work focuses on relatively small scale experiments. Furthermore, we observe that our models are highly tuned to the TED domain and perform poorly on out-of-domain test sets in the considered supervised, as well as unsupervised settings.

Lastly, our setup operates on word-level vocabularies, which presents a computational bottleneck in the output vocabulary. This makes the model difficult to train and deploy in its current implementation. In this work we look into accelerating training using the approach presented by Kumar and Tsvetkov (2019) but we observe a drop in translation quality in doing so. These issues may, however, be addressed with more engineering work, such as using language specific mini-batches (in combination with gradient accumulation), and a dynamic vocabulary sub-selection at training time and at inference time (Senellart et al., 2018).

8 Acknowledgements

References

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *CoRR*, abs/1710.11041.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Stenetorp, Sebastian Riedel, and Mikel Artetxe. 2023. Improving language plasticity via pretraining with active forgetting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations.

Carlos Escolano, Marta R. Costa-jussà, and José A. R. Fonollosa. 2019. From bilingual to multilingual neural machine translation by incremental training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 236–242, Florence, Italy. Association for Computational Linguistics.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, 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.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. 2019. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. 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)*, pages 1890–1902, Minneapolis, Minnesota. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sachin Kumar and Yulia Tsvetkov. 2019. Von mises-fisher loss for training sequence to sequence models with continuous outputs. In *Proc. of ICLR*.

Ruslan Kuprieiev, skshetry, Peter Rowlands, Dmitry Petrov, Paweł Redzyński, Casper da Costa-Luis, David de la Iglesia Castro, Alexander Schepanovski, Ivan Shcheklein, Gao, Batuhan Taskaya, Dave Berenbaum, Jorge Orpinel, Fábio Santos, Daniele, Ronan Lamy, Aman Sharma, Zhanibek Kaimuldenov, Dani Hodovic, Nikita Kodenko, Earl, Andrew Grigorev, Nabanita Dash, George Vyshnya, maykulkarni, Max Hora, Vera, and Sanidhya Mangal. 2024. Dvc: Data version control - git for data & models.

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

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Lori S Levin, Donna Gates, Alon Lavie, and Alex Waibel. 1998. An interlingua based on domain actions for machine translation of task-oriented dialogues. In *ICSLP*, volume 98, pages 1155–1158.

Bryan Li, Mohammad Sadegh Rasooli, Ajay Patel, and Chris Callison-burch. 2023. Multilingual bidirectional unsupervised translation through multilingual finetuning and back-translation. In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 16–31, Dubrovnik, Croatia. Association for Computational Linguistics.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. 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 1259–1273, Online. Association for Computational Linguistics.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020a. On the variance of the adaptive learning rate and beyond. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

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.

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. 2023. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5474–5490, Toronto, Canada. Association for Computational Linguistics.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. 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 244–258, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Pro-*

ceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy. Association for Computational Linguistics.

Jean Senellart, Dakun Zhang, Bo Wang, Guillaume Klein, Jean-Pierre Ramatchandirin, Josep Crego, and Alexander Rush. 2018. OpenNMT system description for WNMT 2018: 800 words/sec on a single-core CPU. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 122–128, Melbourne, Australia. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, 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. In *International Conference on Machine Learning*, pages 5926–5936.

Bernhard Suhm. 1994. Towards better language models for spontaneous speech. In *Proc. ICSLP'94*, volume 2, pages 831–834.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Y Wang and Alex Waibel. 1995. Connectionist transfer in machine translation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria*.

Ye-Yi Wang and Alex Waibel. 1991. A connectionist model for dialog processing. In *icassp*, pages 785–788.

Ye-Yi Wang and Alex Waibel. 1998. Modeling with structures in statistical machine translation. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1357–1363.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

A Vocabulary Details

ℓ	vocab size	CLWE acc.
en	75,234	1.000
de	122,287	0.554
es	93,683	0.750
fr	79,633	0.732
it	96,506	0.696
total	467,343	
ar	203,093	
bg	129,866	0.628
cs	116,719	0.642
ja	54,344	
ko	331,292	
nl	102,027	0.645
pl	164,762	
pt	45,483	0.734
ro	117,749	
ru	178,587	0.627
tr	209,492	
uk	121,020	
vi	48,261	
zh	77,637	

Table 3: The number of words in the vocabulary generated from the training corpus for each of the languages. *total* represents the size of the shared vocabulary of the base model. The column *CLWE acc* describes the nearest neighbour accuracies for the cross-lingual embedding alignments to $\ell_{\mathrm{pivot}} = en$.

B Training Parameters

General Settings	
layers	9
rnn-size	300
word-vec-size	300
transformer-ff	1200
heads	6
warmup-init-lr	1e-8
warmup-end-lr	0.0007
min-lr	1e-9
encoder-type	transformer
decoder-type	transformer
param-init-glorot	True
label-smoothing	0.1
param-init	0
share-embeddings	True
share-decoder-embeddings	True
generator-layer-norm	True
warmup-steps	4000
learning-rate	1
vMF Model	
dropout	0.1
batch-size	40960
batch-type	tokens
normalization	tokens
optim	radam
adam-beta2	0.9995
decay-method	linear
weight-decay	0.00001
max-grad-norm	5.0
lambda-vmf	0.2
generator-function	continuous-linear
loss	nllvmf
Softmax Model	
dropout	0.2
batch-size	5120
batch-type	tokens
accum-count	6
optim	adam
adam-beta2	0.999
decay-method	noam
max-grad-norm	25

Table 4: Training hyperparameter configuration for our main softmax-based model and our main vMF-based model.

C Multilingual Basesystem Scores

We evaluate all our systems using SacreBLEU (Post, 2018). Following the best practice we list the SacreBLEU signature for our BLEU and ChrF++ evalations as follows:
`nrefs:1|case:mixed|eff:no|tok:13a|`
`↪smooth:exp|version:2.4.0`
`nrefs:1|case:mixed|eff:yes|nc:6|nw:2|`
`↪space:no|version:2.4.0`

Table 5 details the ChrF++ scores of our different multilingual models averaged over their 20 supervised translation directions. Table 6 and 7 detail the individual scores.

model	ChrF++	Δ
BPE baseline	55.3	
softmax CLWE	55.1	-0.16
vMF CLWE	51.5	-3.79

Table 5: Comparison of average ChrF++ scores on the 20 supervised directions for the different variants of the multilingual base system.

D Supplementary Zero-Shot Translation Scores

Table 8 details additional zero-shot translation scores, for languages and translation directions not listed in Table 1. The scores are produced with our main Romance-Germanic model, so the seen languages are English, German, Spanish, French and Italian and model dimension $d_{\mathrm{model}} = 300$.

E Unsupervised MT BLEU Scores

Table 9 details the BLEU scores for the Table 2 evaluations.

F Effect of Domain

To study the effect of domain and amount of data we perform iterative back-translation training on Wikipedia and news domain. For domain specific test data we use Flores, to evaluate the Wikipedia domain, and wmt16 test data for evaluation of the news domain. We detail the results in Table 10. When trained on Wikipedia we observe a slight increase in ChrF++ scores on English-Portuguese Wikipedia domain, but a large drop on the original TED domain. In all other cases we observe drops in performance on all considered domains.

G Estimate of GPU Hours

We conduct our preliminary experiments on an `Nvidia GTX 1080Ti` GPU. The main body of our experiments if conducted on a cluster of `Nvidia Titan RTX` GPUs. Our largest softmax-based full-word vocabulary model (with a count of 53M trainable parameters, 317M non-trainable) is trained on an `Nvidia A6000` GPU. The GPU hours roughly sum up to 1039 GPU hours, 197h out of these on an `Nvidia A6000` GPU.

ℓ	en	de	es	fr	it
en		54.1	61.4	61.0	57.9
de	57.5		50.3	52.1	48.7
es	62.6	49.0		56.4	54.2
fr	61.1	49.2	54.9		53.5
it	59.7	48.4	54.8	55.7	

ChrF++

ℓ	en	de	es	fr	it
en		30.0	39.3	44.0	36.5
de	38.8		26.5	33.6	26.1
es	44.5	24.3		38.4	32.0
fr	43.1	24.8	31.5		31.7
it	41.2	23.9	31.6	37.6	

BLEU

Table 6: ChrF++/BLEU scores of our main softmax-based model on the TED test split (Qi et al., 2018). The reported languages present the supervised training directions. Rows represent the source language, while columns represent the target language.

ℓ	en	de	es	fr	it
en		54.0	61.5	61.2	58.1
de	57.9		50.5	52.6	49.0
es	62.7	48.8		56.6	54.3
fr	61.4	49.4	55		53.7
it	59.8	48.4	54.8	55.9	

ChrF++

ℓ	en	de	es	fr	it
en		29.6	39.1	43.9	36.4
de	39.0		26.7	33.8	26.1
es	44.4	24.0		38.4	32.1
fr	43.3	24.9	31.5		31.6
it	41.3	23.9	31.5	37.6	

BLEU

Table 7: ChrF++/BLEU scores of our BPE-based baseline model on the TED test split (Qi et al., 2018). The reported languages present the supervised training directions. Rows represent the source language, while columns represent the target language.

ℓ	en	de	es	fr	it
pt	61.0	48.4	57.1	56.2	53.6
ru	42.1	38.0	39.8	41.5	38.5
pl	40.5	36.1	38.2	39.9	36.8
uk	41.3	36.5	38.9	40.1	37.6
zh	33.3	29.0	31.2	32.1	29.7
vi	34.6	30.4	32.5	32.7	31.1
ja	28.3	26.0	27.2	27.8	26.2
ko	25.0	23.2	24.4	24.5	23.0
ar	38.3	32.1	35.8	36.3	34.0
tr	31.9	28.9	31.1	31.0	29.8
nl	51.3	45.5	46.4	48.3	44.5
bg	48.9	42.6	46.6	47.3	44.1
cs	45.2	40.0	42.2	44.2	40.9
ro	49.0	42.1	46.8	48.1	44.8

ChrF++

ℓ	en	de	es	fr	it
pt	42.4	23.2	34.0	38.1	31.1
ru	20.4	13.3	14.9	21.6	15.0
pl	20.0	12.1	14.4	20.5	14.0
uk	20.0	11.7	14.1	20.0	14.2
zh	9.9	4.1	5.3	10.4	5.6
vi	12.4	5.9	8.0	12.1	7.9
ja	4.4	1.3	2.2	5.7	2.5
ko	5.4	1.6	2.6	5.8	2.7
ar	17.0	7.7	10.9	16.2	10.5
tr	9.2	5.0	6.9	10.4	7.1
nl	31.0	20.4	22.7	29.6	21.8
bg	28.0	17.2	21.6	27.8	20.3
cs	23.1	15.1	17.5	24.6	17.1
ro	28.1	16.9	22.3	28.6	21.0

BLEU

Table 8: ChrF++/BLEU scores of our main softmax-based model on the TED test split (Qi et al., 2018). The reported languages present translations from unseen source languages. Rows represent the source language, while columns represent the target language. Note that the translations are produced with a beam size of 1 and thus slightly differ from the scores reported in Table 1.

Model	en→pt			en→ru			en→tr			en→ko		
	TED	Flores	Tatoeba	TED	Flores	Tatoeba	TED	Flores	Tatoeba	TED	Flores	Tatoeba
(Song et al., 2019)							4.0	4.8	1.3			
(Artetxe et al., 2019)							0.4	0.2	0.3			
Iteration 1	36.8	27.4	33.4	17.4	14.1	25.2	8.1	5.4	8.2	8.0	7.0	7.9
Iteration 3	37.4	28.0	33.7	18.5	15.7	27.5	11.0	6.9	13.8	11.4	8.6	9.8
Iteration 5	37.3	27.9	33.7	18.7	16.4	29.9	12.5	7.6	16.5	12.5	9.4	10.8
Supervised	36.4	26.8	32.2	20.1	16.5	29.6	15.7	10.2	21.1	14.0	11.2	13.8

Model	pt→en			ru→en			tr→en			ko→en		
	TED	Flores	Tatoeba	TED	Flores	Tatoeba	TED	Flores	Tatoeba	TED	Flores	Tatoeba
(Song et al., 2019)							4.1	5.9	2.9			
(Artetxe et al., 2019)							0.2	0.1	0.1			
Zero-shot	42.6	30.6	38.7	20.7	13.8	24.8	9.5	5.0	6.4	5.6	2.5	3.0
Iteration 2	44.0	31.1	40.6	24.9	17.0	31.8	15.6	9.3	17.9	9.6	4.1	7.2
Iteration 4	43.6	30.8	39.9	25.9	18.5	35.6	19.0	10.8	24.5	10.5	4.4	8.0
Iteration 6	43.5	30.8	39.5	26.0	18.7	36.2	21.1	12.1	28.6	10.8	4.6	8.4
Supervised	45.2	31.0	41.3	28.3	19.2	39.4	27.7	17.5	35.7	19.3	9.6	21.8
# of Sentences	52k			208k			182k			206k		

Table 9: BLEU scores after each iteration of back-translation (see Table 2). In each iteration we use our Romance-Germanic multilingual model to zero-shot translate from TED domain to create synthetic parallel data.

	train on	#sentences	TED	Flores	wmt16
en→tr	TED	182k	41.1	38.7	
	Wiki-40B	1947k	32.8	37.8	
en→pt	TED	51k	57.3	50.5	
	Wiki-40B	6094k	51.6	51.8	
en→ro	TED	180k	49.1	48.0	43.2
	NewsCrawl	2281k	44.2	44.9	43.9

Table 10: A comparison of average test scores for different variations of the autoencoder methods (top), as well as the back-translation method (bottom).