

# Unsupervised Transfer Learning in Multilingual Neural Machine Translation with Cross-Lingual Word Embeddings

Carlos Mullov and Ngoc-Quan Pham and Alexander Waibel

Institute for Anthropomatics and Robotics  
KIT - Karlsruhe Institute of Technology, Germany  
{firstname}.{lastname}@kit.edu

## Abstract

In this work we look into adding a new language to a multilingual NMT system in an unsupervised fashion. Under the utilization of pre-trained cross-lingual word embeddings we seek to exploit a language independent multilingual sentence representation to easily generalize to a new language. While using cross-lingual embeddings for word lookup we decode from a yet entirely unseen source language in a process we call *blind decoding*. Blindly decoding from Portuguese using a basesystem containing several Romance languages we achieve scores of 36.4 BLEU for Portuguese-English and 12.8 BLEU for Russian-English. In an attempt to train the mapping from the encoder sentence representation to a new target language we use our model as an autoencoder. Merely training to translate from Portuguese to Portuguese while freezing the encoder we achieve 26 BLEU on English-Portuguese, and up to 28 BLEU when adding artificial noise to the input. Lastly we explore a more practical adaptation approach through non-iterative backtranslation, exploiting our model's ability to produce high quality translations through blind decoding. This yields us up to 34.6 BLEU on English-Portuguese, attaining near parity with a model adapted on real bilingual data.

## 1 Introduction

Neural machine translation (NMT) systems consist of an encoder, which encodes an input sentence into a latent numerical sentence representation space, and a decoder, which generates the target sentence from this representation (Cho et al., 2014; Bahdanau et al., 2015). While under favourable circumstances neural networks bring significant improvements to translation performance, a major problem with them, however, is that they are extremely data hungry. While for the most common language pairs large amounts of parallel data might be available, for most language pairs parallel data is extremely

scarce. As such one of the goals of multilingual NMT is to significantly reduce the amount of parallel data required for each given individual language pair. Universal multilingual NMT (Johnson et al., 2017; Ha et al., 2016) employs a universal encoder so that the source languages share one single set of parameters.

Ideally this universal encoder would learn a language independent representation of the input sentences, e.g. learn to map semantically similar sentences onto similar neural sentence representations – even across different languages. Ongoing research in the field of zero-shot translation provided evidence that multilingual NMT models exhibit this property up to a certain extent, and that enforcing the similarity between sentence representations across different languages in turn also improves the zero-shot capabilities (Pham et al., 2019). On the other hand, cross-lingual transfer learning research in NMT further showed that such multilingual systems can rapidly be extended to new languages on very little data (Neubig and Hu, 2018). Neubig and Hu (2018) furthermore show that going as far as training a basesystem on 58 source languages immensely helps the universal encoder in subsequently generalizing to new source languages in a supervised setting. In this work we take this approach one step further and investigate the ability to translate to a new language with no parallel data for this language. While integrating cross-lingual word embeddings into the model we provide two contributions in the field of multilingual NMT:

1. To explore the generalization-ability of the universal encoder we translate from a yet entirely unseen language to a known language.
2. Inspired by the recent advancements in unsupervised NMT we adapt the universal decoder to a new language through
  - a) using the NMT model as autoencoder
  - b) backtranslation

Using a German-English-Spanish-French-Italian multilingual basystem we perform experiments in transfer learning to translate to and from Portuguese and Russian. First, by simply decoding from Portuguese and Russian while using the cross-lingual embeddings for word lookup we demonstrate the effectiveness of applying the universal encoder to an unseen language, achieving reasonable 36.4 BLEU and 12.8 BLEU scores for Portuguese-English and Russian-English respectively. We discovered that, when using the multilingual model as autoencoder, we can successfully adapt to the new language even without training to denoise – simply by freezing the encoder in training. Finally, we devise a method to jump-start the iterative backtranslation process oftentimes employed in unsupervised NMT, by exploiting model transfer learning ability and generate data by translating from an unseen language. Our models trained on the synthetic parallel data attain translation performance close to what we achieve through supervised adaptation on real parallel data, falling short by an average of 0.8 BLEU for translation to Portuguese.

As our final contribution we experiment on continuous output NMT, which, while underperforming the softmax models in terms of BLEU scores, attains vastly superior performance in terms of training times as well as memory complexity.

## 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  $\ell_{tgt}$  is a suitable translation of the source sentence  $x$  in the source language  $\ell_{src}$ . We describe the distribution of the sentences in different natural languages  $\ell$  through random variables  $X_\ell, Y_\ell$ . A universal encoder maps the sentences from the different input distributions onto a single, shared latent distribution  $H_{enc} = enc(X_\ell)$ . A decoder is then tasked to model the probability distribution from this latent variable:  $\mathbb{P}(Y_{\ell_{tgt}} = y \mid H_{enc} = c)$ . 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 log-likelihood of a dataset of parallel sentences  $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ , where sen-

tence 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 the only difference to bilingual translation is the nature of the input and output distribution. Thus, the only 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.

It is a desirable quality for the universal encoder to learn a language independent latent representation. Such language independence may be expressed as a low variance  $\mathbb{V}\{H_{enc}\}$  in the latent encoding space, as well as a high cross-lingual covariance  $\mathbb{C}\{enc(X_\ell), enc(X_{\ell'})\}$ .

## 3 Related Work

Qi et al. (2018) looked into the effectiveness of using pre-trained embeddings for NMT, asking whether the alignment of the embedding vectors into a shared embedding space helps in NMT and coming to the conclusion that it is helpful in a multilingual setting. Similarly, Sen et al. (2019) carried out unsupervised multilingual NMT with the utilization of cross-lingual word embeddings. While their approach is similar to ours, their focus lies in improving unsupervised NMT, while we mainly look into multilingual transfer learning. 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 came to the conclusion that pre-training in a highly multilingual setting – this in their case is a system with 58 source languages – significantly improves the ability to learn the low-resourced language. Similar to our findings, they also achieved significant performance on a yet entirely unseen language. They, however, do not use cross-lingual embeddings, but rely on bilingual data to teach their model the cross-lingual word correlations. Kim et al. (2019) conducted unsupervised cross-lingual transfer learning in NMT, transferring the parameters learned on a high resource language pair and to a new source language. Similar to our work they therefore employed cross-lingual word embeddings, aligning the new language embeddings to the parent model source embeddings. Their work

focuses on adaptation on the source-side, furthermore in a bilingual setting. Escolano et al. (2019) devised an approach to multilingual NMT with independent encoders and decoders that allows for zero-shot translation, as well as the addition of new languages. Siddhant et al. (2020) looked into leveraging monolingual data for multilingual NMT. Amongst other things they extended the model by a new language through masked denoising autoencoding. While they did not perform backtranslation they suggest that their scenario presents a promising avenue for jump-starting the backtranslation process, which our work demonstrates to indeed be the case.

Our work is also closely related to unsupervised NMT (Lample et al., 2017, 2018; Artetxe et al., 2017, 2019). Lample et al. (2017) trained an unsupervised NMT system which translates between two languages  $l_1$  and  $l_2$ , while employing universal encoder and decoder for the simultaneous translation in both directions. They iteratively trained the NMT model to generate  $l_1$  and  $l_2$  sentences through denoising autoencoding, while simultaneously training to maximize the similarity between the encoder sentence representation for  $l_1$  and  $l_2$  sentences. Through unsupervised means they thus induce a shared sentence representation, essentially bootstrapping a multilingual NMT system from monolingual data. We, on the other hand, consider a more realistic scenario where we have enough data from high resource language pairs to train a supervised multilingual NMT system, on which we then perform unsupervised transfer learning.

## 4 Approach

We propose a three-step approach to the cross-lingual transfer learning task

1. train the cross-lingual word embeddings on monolingual data for each of the involved languages
2. train the multilingual basesystem on parallel data for the set of base languages  $\mathcal{L}_{\text{base}}$
3. add a new language  $l_{\text{new}} \notin \mathcal{L}_{\text{base}}$  to the system on monolingual data only

Figure 1 describes the resulting NMT architecture.

### 4.1 Cross-Lingual Word Embeddings

In conventional training of a universal multilingual NMT system the  $l_{\text{new}}$  word vectors are randomly initialized. The model then learns in training to

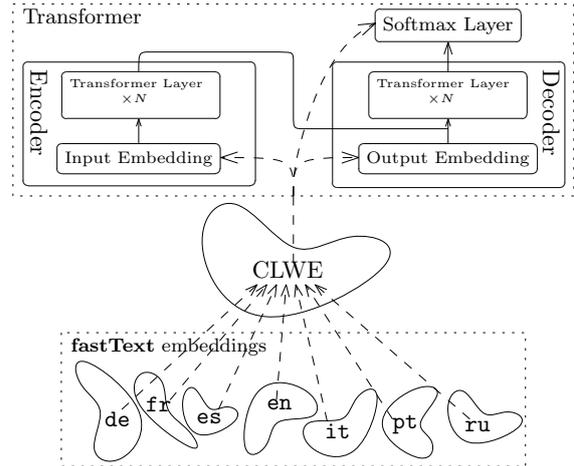


Figure 1: Our NMT architecture consists of a Transformer model with pre-trained cross-lingual word embeddings for embedding layers. The embedding vectors are obtained through alignment of monolingual *fastText* embeddings for each language into a common embedding space.

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 at achieving an equivalent result through cross-lingual word embeddings. These word embeddings are then integrated into our translation model by supplying in the form of pre-trained embeddings.

For each language  $l \in \mathcal{L}_{\text{base}} \cup \{l_{\text{new}}\}$  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 we choose to align into a *single hub space* (SHS Heyman et al., 2019), e.g. we pick one of our base languages as a *pivot* and align each of the embedding spaces  $E_l$  to  $E_{\text{pivot}}$ .

### 4.2 Adding a New Language

In this work we present three methods to add a new language to a multilingual basesystem. In the order of increasing level of supervision, they are as follows:

1. *blind* decoding
2. (denoising) autoencoding
3. backtranslation

The first experiment serves to answer the question of how well a pre-trained multilingual model can generalize to an unseen language, and can equally be applied on the source and on the target-side.

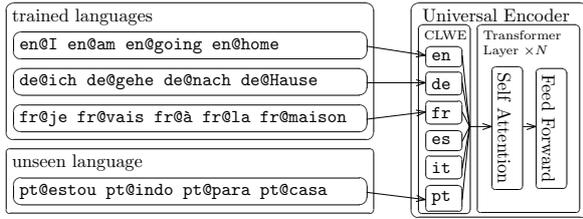


Figure 2: In *blind decoding* we simply decode from an unseen language  $\ell_{\text{new}}$  without prior exposure to  $\ell_{\text{new}}$  whatsoever. The cross-lingual word embeddings (CLWE) serve as lookup table for  $\ell_{\text{new}}$  words.

The latter two experiments present two different methods for adaptation on  $\ell_{\text{new}}$  monolingual data, and focus on adding  $\ell_{\text{new}}$  on the target-side.

#### 4.2.1 Blind Decoding

To provide an indication of how well the universal encoder generalizes to unseen languages we let our multilingual model decode from  $\ell_{\text{new}}$  sentences, without any sorts of prior exposure to  $\ell_{\text{new}}$ . We call this method *blind* decoding. As depicted in Figure 2 we therefore employ the cross-lingual embeddings to lookup the  $\ell_{\text{new}}$  word vectors and then simply decode from the resulting sequence. While we are also concerned about testing the NMT system ability to blindly decode *to* a new language, we consider the target-side the less interesting side in this experiment, as the decoder couldn’t possibly know the sentence level features, such as the  $\ell_{\text{new}}$  word order. We anticipated that the language production ability should be very limited, while the language comprehension ability should – through the information supplied by the word embeddings – only be constrained to a lesser extent. While the encoder may not know the syntax of an unseen language either, many syntactical concepts in language – such as grammatical case, part-of-speech, grammatical genders, or tense – are encoded at the word level<sup>1</sup>. 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.

#### 4.2.2 Learning Sentence-Level Features via Autoencoding

As our first adaptation method towards  $\ell_{\text{new}}$  we propose two different ways of using our NMT model as an autoencoder, e.g. train to translate from  $\ell_{\text{new}}$

<sup>1</sup>Take German for example, a noun would have different surface forms for different grammatical cases, which the fast-Text model should take into account and appropriately assign different word vectors.

to  $\ell_{\text{new}}$ . The primary objective in doing so is to train the decoder  $\ell_{\text{new}}$  language model while maintaining the cross-lingual signal received from the encoder.

**Denosing autoencoder** For the first autoencoding approach we want to focus on teaching the model the target language syntax. To achieve this, we train the model to perform denosing autoencoding: the model gets noisy input sentences and is tasked with reconstructing the original, grammatically correct version of this sentence. One major aspect of language syntax which sets apart different languages from each other is the word order. By “slightly” shuffling the input sentence (Lample et al., 2017), e.g. displacing every word by up to  $n$  positions from the original position, we can easily force the model to learn the correct word order.

**Plain autoencoder** As our second autoencoding<sup>2</sup> method we try training raw  $\ell_{\text{new}} \rightarrow \ell_{\text{new}}$  translation while freezing the encoder parameters. With this method we rely on the ability of our encoder to encode sentences from  $\ell_{\text{new}}$  in a suitable manner, as we directly take the encoder output for  $\ell_{\text{new}}$  sentences and learn to generate  $\ell_{\text{new}}$  sentences from it. In order for the model to not just learn to copy words from the input sentence to the output sentence, we rely on the encoder output to be language independent enough to not retain any word level information from the source sentence. This method thus tests the raw ability of our NMT system to learn the  $\ell_{\text{new}}$  language generation task conditioned on the encoder output  $c$ :  $\mathbb{P}(Y_{\ell_{\text{new}}} = y \mid H_{\text{enc}} = c)$ . In worst case this method trains the language generation ability conditioned on white noise.

#### 4.2.3 Backtranslation

Lastly we would like to combine our model’s ability to translate from  $\ell_{\text{new}}$  and its ability to adapt to  $\ell_{\text{new}}$ . We therefore use the in unsupervised NMT commonly employed approach of training on *backtranslated* data (Sennrich et al., 2016). Therefore one can use a joint NMT system translating between two languages in both directions to iteratively produce better translations, to then adapt the bootstrapped model on the improved synthetic generated data (Lample et al., 2017). In our approach to backtranslation – exploiting the model’s ability to perform blind decoding – we choose a simplified approach: we perform one single round of

<sup>2</sup>due to the differing nature of the task we consider *autodecoding* might be a more appropriate name

lng	vocab size	CLWE acc
en	47,067	1.000
de	98,747	0.554
es	74,267	0.750
fr	60,416	0.732
it	74,962	0.696
total	355,459	
pt	67,237	0.734
ru	166,651	0.627

Table 1: 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  $\ell_{\text{pivot}} = \text{en}$ .

backtranslation in which we use monolingual  $\ell_{\text{new}}$  data to 1) blindly decode from  $\ell_{\text{new}}$  to each one of our base languages  $\ell \in \mathcal{L}_{\text{base}}$  in order to generate synthetic parallel data, and then 2) reversing the translation direction to train our model on this synthetic parallel data to translate from  $\ell$  to  $\ell_{\text{new}}$ .

## 5 General Experimental Setup

**Cross-lingual word embeddings** As the basis for all of our monolingual word embeddings we use the pre-trained *fastText* models provided on the fastText website<sup>3</sup>. These models provide 300-dimensional high-quality embeddings for full word units, accumulated on large web crawled monolingual text corpora. Using the pivot approach described in section 4.1 we align these embeddings into a common embedding space using direct supervised optimization on the RCSSL criterion (Joulin et al., 2018) while using the bilingual dictionaries provided by the MUSE repository<sup>4</sup>. In order to reduce the vocabulary size, we subsequently regenerate our vocabulary and the corresponding embedding vectors using only the words in our NMT training corpus. Alignment accuracies and vocabulary sizes are listed in Table 1.

**Full word-units** While using a joint BPE lets us share information between vocabularies and significantly reduces the size of the multilingual vocabulary, for our approach we choose to work with full word units, since our preliminary experiments

have shown them to outperform the BPE models when used in combination with cross-lingual embeddings. Furthermore, this gives us better compatibility with distant languages with separate alphabets and to avoid 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.

For our multilingual vocabulary 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*. This also allows us to easily match any token in the vocabulary to its corresponding language, which is not trivial in a model with BPE units that are shared across multiple languages. In inference we thus restrict the target vocabulary to the tokens matching the language we decode to, thereby forcing the model to only choose tokens from a single language in a situation where it would otherwise not do so, e.g. to decode to a yet entirely unseen language (see section 6.1). This filtering furthermore provides a speedup in inference, as well as an improvement by around +0.4 BLEU in our basystem evaluation.

**Datasets** As our training, development and evaluation data sets we use multilingual transcriptions of TED talks (Cettolo et al., 2012). For our basystems we include English, German, Spanish, French and Italian as the base languages. We train the basystem on a total of 20 language pairs, and a parallel corpus size of 3, 251, 582 sentences. For our transfer learning experiments we use monolingual TED data for Portuguese and Russian, and parallel test data for BLEU score evaluation. The monolingual data corpora contain 148, 321 sentences of Portuguese data and 187, 843 sentences of Russian data. For test and development data we use the IWSLT dev2010 and tst2010 data sets. All data is first tokenized and truecased with standard Moses tools.

**NMT model** For our NMT model we use a Transformer model (Vaswani et al., 2017) with relative position encodings (Shaw et al., 2018). In accordance with our 300-dimensional fastText embeddings we use a model size of 300, and inner size

<sup>3</sup><https://fasttext.cc/docs/en/crawl-vectors.html>

<sup>4</sup>(Conneau et al., 2017, <https://github.com/facebookresearch/MUSE>)

of 1200, 6 attention heads and 6 layers. Our implementation is based on the repository provided by Kumar and Tsvetkov (2018)<sup>5</sup>, which in turn is based on *OpenNMT-py* (Klein et al., 2017). We share the multilingual vocabulary, as well as the embedding parameters across the encoder and the decoder, and use tied output embeddings. Refer to Appendix A for the full set of training parameters.

The evaluation scores for the multilingual base-model, which we train on our available parallel data are listed in Table 2. The scores represent the average over all of the 20 language pairs. For the individual scores refer to Appendix B. The *frozen softmax* model serves as the main model for our transfer learning experiments. It uses the cross-lingual embeddings in a pure lookup fashion, never updating the word vectors in NMT training. This freezing ensures that the word embeddings stay in the shared embedding space, and we can as such trivially add new words to our vocabulary and align new languages into the shared embedding space. The regular non-frozen softmax model serves for comparison with the frozen softmax model, as we do not know how well the pre-trained fastText embeddings fit the translation task. We initialize this model to the trained cross-lingual embeddings while in training adapting them to the training data. We can see that the frozen softmax model performs on par with its non-frozen counterpart, which suggests that the cross-lingual embeddings are very well suited for the translation task, despite them being trained on an unrelated task on purely monolingual data.

For the *continuous output* model we employ the approach proposed by Kumar and Tsvetkov (2018), which, instead of calculating a multinomial distribution over a closed vocabulary, optimizes the network output directly towards the semantic information encoded in our pre-trained word embeddings. This is ideal for our use case, since this is more suited towards extending the vocabulary. Furthermore it helps us deal with our large multilingual vocabulary, as computational complexity and memory complexity in a continuous output model are independent of its vocabulary size. While the model delivers inferior performance considering raw BLEU scores only, training times amount to around one fifth of the training times for the softmax models.

<sup>5</sup><https://github.com/Sachin19/seq2seq-con>

model	dev	$\Delta$	test	$\Delta$
BPE baseline	26.10		27.73	
softmax	24.56	-1.54	26.65	-1.08
frozen softmax	24.96	-1.14	26.60	-1.13
continuous	22.27	-3.83	24.11	-3.62

Table 2: Comparison of average BLEU scores for the different variants of the multilingual base system

Finally, in order to test the viability of our basesystem – namely a full word-unit translation system with pre-trained word embeddings – we also train a standard subword-unit translation system. This model uses BPE trained on the multilingual corpus, reducing the vocabulary size from the 355, 459 words in our full word-unit vocabulary to 36, 150 BPE units.

## 6 Adding a New Language

In the following we present our results for extending the multilingual basesystem by Portuguese and Russian. Unless stated otherwise, we evaluate our methods on the section 5 *frozen softmax* model. For our adaptation methods we select the best performing model based on the development data BLEU scores.

### 6.1 Blind Decoding

As our first experiment we investigate decoding to  $\ell_{\text{new}}$  without any sort of additional training or adaptation. We therefore simply swap out the source or target-side embeddings and decode from or to  $\ell_{\text{new}}$ . As seen in Table 3 we can achieve BLEU scores of up to 36.4 with Portuguese as the source language. While the average of 28.3 BLEU on the test set is significantly lower, it is still a considerably reasonable score, considering the model has never seen even a single Portuguese sentence. On a more distant source language such as Russian, we achieve much lower scores. Despite of the fact that the basesystem has never even seen a single sentence from any Slavic language the results are, however, still intelligible, achieving scores of up to 12.8 BLEU.

When decoding with  $\ell_{\text{new}}$  on the target-side, to force the model to decode to  $\ell_{\text{new}}$  as opposed to any  $\ell \in \mathcal{L}_{\text{base}}$  we need to filter out any target vocabulary words not associated with  $\ell_{\text{new}}$ . Initially, using our frozen softmax model to decode to Portuguese results in an average BLEU score of 3.0 BLEU. For

source-side method	pt-de	pt-en	pt-es	pt-fr	pt-it	ru-de	ru-en	∅ pt	∅ ru
blind decoding	20.6	36.4	30.7	29.3	24.4	11.1	12.8	28.3	11.9
supervised	22.2	39.7	33.0	31.9	26.6	15.2	20.9	30.7	18.1
target-side method	de-pt	en-pt	es-pt	fr-pt	it-pt	de-ru	en-ru	∅ pt	∅ ru
blind decoding	8.4	16.2	13.3	11.6	10.7	1.1	1.7	12.0	1.4
autoencoder	17.0	28.1	27.1	21.5	20.8	8.1	8.7	22.9	8.4
backtranslation	21.3	34.6	32.3	26.4	25.4	13.6	13.9	28.0	13.8
supervised	21.9	35.8	32.9	27.1	26.2	15.1	17.2	28.8	16.2

Table 3: BLEU scores for decoding from (top) and to (bottom) Portuguese and Russian as the new language, with and without adaptation through various methods.

the continuous output model this amounts to 6.2 BLEU. We suspect the main cause for the inferior softmax performance to be the missing bias for the new language, which the continuous output system does not require. In fact, the models often get stuck in a decoding loop, which we combat by adding a postprocessing step remove the resulting duplicate words. For both models this doubles the BLEU score, achieving up to even 16.2 BLEU on English-Portuguese.

## 6.2 (Denoising) Autoencoder

For our autoencoding methods we adapt on monolingual TED data using unchanged training parameters from the basystem training. When training to denoise via word order shuffling, we found – in accordance with [Lample et al. \(2017\)](#) – that displacing every word by up to  $n = 3$  positions yields the best results. We furthermore experiment with word dropout as additional noise, however, we did not observe any noticeable improvement over pure reordering noise. Table 4 shows the results for afterwards decoding to Portuguese and to Russian. The denoising autoencoding results in slightly lower scores than the frozen encoder variant. While the resulting scores clearly demonstrate the model ability to adapt to the new language – even without noisy input – we find the Russian scores to be disappointingly low. Using both methods in conjunction with our postprocessing step (see section 6.1), we bring the translation to Russian up to a maximum of 8.7 BLEU and translation to Portuguese up to a maximum of 28.1 BLEU. Table 3 shows the resulting BLEU scores for the individual language pairs. The Portuguese and Russian softmax models both train for 1,000 iterations, completing after roughly 60 minutes on a *GTX 1080 Ti* GPU.

We believe the significant improvement for the

frozen encoder approach when adding noise to be an indication that the encoder representation still retains certain language specific word-level information. In this case the noiseless model might find it easier to reconstruct the original sentence, thus learning less about the new language syntax in the process. While this is an undesirable quality, this is to be expected considering just the fact that the encoder output sequence is equal in length to the input sentence.

## 6.3 Backtranslation

Without any additional steps, our blind decoding approach has already enabled our basystem to produce translations of up to 36.4 BLEU from Portuguese. In our final experiment we exploit this for backtranslation to benefit the more difficult direction. We train the Portuguese and the Russian model on backtranslated data for 6,000 and 10,000 iterations respectively, again with unchanged training parameters from basystem training. Table 3 shows the translation scores after training on backtranslated data for each of the language pairs.

method	∅ pt	∅ ru
denoising autoencoder	20.8	6.5
frozen encoder	21.1	6.6
+ denoise	22.9	7.9
+ postprocessing	22.9	8.4
backtranslation	27.4	13.2
+ frozen encoder	28.0	13.8
supervised baseline	28.8	16.2

Table 4: A comparison of average test scores for different variations of the autoencoder methods (top), as well as the backtranslation method (bottom).

Translation scores reach up to 34.6 BLEU, namely on English-Portuguese. From the comparison in Table 4 we can also see that the Portuguese backtranslation scores are almost on par with our supervised model trained on bilingual data, falling short by just 0.8 BLEU. Again, freezing the encoder in training results in improved translation, albeit to a lesser extent than for the autoencoding approach. We note that training the Russian system on  $en \rightarrow ru$  data only also improves the performance for  $de \rightarrow ru$ , in our experiments sometimes even outperforming the language pair it is trained on. We see this as an indication that this training mainly affects the decoder language model for Russian. This furthermore means that the decoder, when decoding to Russian, is able to use a similar amount of information from the encoder representation of the source sentence. This either means that the encoder representation of English and German are very similar in quality, being able to extract the same amount of meaning. Alternatively it means that the decoder does not learn to use the full amount of information encoded into the representation of the English source sentence, meaning the decoder is lacking in translation adequacy while leaning towards better fluency.

## 7 Conclusion

In this work we have looked into adding a new language to a previously trained multilingual NMT system in an unsupervised fashion. We explored the possibility of reusing the existing latent sentence representation and adding a language by adapting the decoder on monolingual data only. As part of our approach we show the feasibility of doing multilingual NMT with pre-trained cross-lingual word embeddings. By blindly decoding from Portuguese, e.g. translating without any prior exposure towards Portuguese aside from the cross-lingual word embeddings, using a basesystem containing multiple Romance languages we achieve scores of up to 36.4 BLEU for Portuguese-English, thereby demonstrating the universal encoder ability to generalize to an entirely unseen language. While this works significantly worse on a more distant language, this model – which has never seen even a single sentence from any Slavic language – is still able to achieve up to 13 BLEU for Russian-English. Furthermore, by applying blind decoding on the target-side we have been able to achieve up to 16.2 BLEU when decoding *to* Portuguese. To this end

we looked into the recently proposed continuous output representation approach.

In an attempt to train the mapping from our sentence representation to a new target language we use our model as an autoencoder. By training to translate from Portuguese to Portuguese while freezing the encoder we have achieved up to 26 BLEU for English-Portuguese. Adding artificial noise to the input to let the model learn the correct word order gains us an additional 2 BLEU. Translating to Russian we achieve up to 8.7 BLEU for English-Russian.

Lastly we have explored a more practical approach of learning the new language by training the system on backtranslated data. To this end we exploit our model’s ability to produce high quality translations from an unseen source-side language to generate the synthetic data. The training on the synthetic data has yielded up to 34.6 BLEU, again on English-Portuguese, attaining near parity with a model trained on real bilingual data. Translating to Russian yields at most 13.9 BLEU for English-Russian. Considering the low English-Russian baseline score of 17.2 BLEU we suspect the overall low Russian scores to partly be an issue with the low quality word embedding alignment.

### 7.1 Future Work

In the future we would like to explore the question of how the composition and number of languages in the base system affects the ability to perform the transfer learning, and whether seeing a wide variety of different languages will help the translation system in generalizing to new languages – especially more distant ones. In order to further improve the performance on source-side unseen languages we would like to explore various methods for the adaptation of the encoder. While the focus of this work lies mainly on adapting the decoder, an improvement on the source-side would especially benefit the backtranslation results for more distant languages. As an alternative to teaching the new language to the decoder via autoencoding, we would like to explore the possibility of using generative adversarial training on top of a translation system with a continuous output representation.

## 8 Acknowledgement

The projects on which this paper is based were funded by the Federal Ministry of Education and Research (BMBF) of Germany under the numbers

01IS18040A. The authors are responsible for the content of this publication.

## References

- 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.
- 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.
- Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. [Wit<sup>3</sup>: Web inventory of transcribed and translated talks](#). In *Proceedings of the 16<sup>th</sup> Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using RNN encoder–decoder for statistical machine translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. [Word translation without parallel data](#). *arXiv preprint arXiv:1710.04087*.
- 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.
- 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.
- 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.
- 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. 2018. [Von mises-fisher loss for training sequence to sequence models with continuous outputs](#). *CoRR*, abs/1812.04616.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2017. [Unsupervised machine translation using monolingual corpora only](#). *arXiv preprint arXiv:1711.00043*.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018. [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.
- 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.

- Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. [Improving zero-shot translation with language-independent constraints](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. 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.
- 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.
- 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.

## A Training Parameters

general settings:

```
layers 6
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
position_encoding
max_generator_batches 2
param_init_glorot
label_smoothing 0.1
param_init 0
share_embeddings
share_decoder_embeddings
generator_layer_norm
warmup_steps 4000
learning_rate 1
```

vmf model:

```
dropout 0.1
batch_size 8192
batch_type tokens
normalization tokens
accum_count 2
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 1536
batch_type tokens
accum_count 6
optim adam
adam_beta2 0.999
decay_method noam
max_grad_norm 25
```

## B Multilingual Basesystem Scores

$\ell$	de	en	es	fr	it
de		30.6	21.4	23.0	18.5
en	25.9		34.6	33.1	26.5
es	21.8	40.8		29.8	25.1
fr	19.1	32.5	25.3		22.8
it	18.9	30.2	25.7	26.7	

Table 5: Individual language pair test scores for the section 5 *frozen softmax* basesystem. Lines represent the source language, while columns represent the target language.