# Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies

**Yunsu Kim** **Yingbo Gao** **Hermann Ney**
Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
{surname}@cs.rwth-aachen.de



## Abstract

Transfer learning or multilingual model is essential for low-resource neural machine translation (NMT), but the applicability is limited to cognate languages by sharing their vocabularies. This paper shows effective techniques to transfer a pre-trained NMT model to a new, unrelated language without shared vocabularies. We relieve the vocabulary mismatch by using cross-lingual word embedding, train a more language-agnostic encoder by injecting artificial noises, and generate synthetic data easily from the pre-training data without back-translation. Our methods do not require restructuring the vocabulary or retraining the model. We improve plain NMT transfer by up to +5.1% BLEU in five low-resource translation tasks, outperforming multilingual joint training by a large margin. We also provide extensive ablation studies on pre-trained embedding, synthetic data, vocabulary size, and parameter freezing for a better understanding of NMT transfer.

## 1 Introduction

Despite recent success of neural machine translation (NMT) (Bahdanau et al., 2015; Vaswani et al., 2017), its major improvements and optimizations cannot be easily applied to low-resource language pairs. Basic training procedure of NMT does not function well with only a handful of bilingual data (Koehn and Knowles, 2017), while collecting bilingual resource is arduous for many languages.

Multilingual NMT solves the problem of lacking bilingual data by training a shared model along with other related languages (Firat et al., 2016; Johnson et al., 2017). For this to work in practice, however, we need a considerable effort to gather bilingual data over multiple languages and preprocess them jointly before training. This has two critical issues: 1) The languages for training should be linguistically related in order to build a shared vocabulary. 2) It is not feasible to add a new language to a trained model, since the training vocabulary must be redefined; one may need to re-train the model from scratch.

In transfer learning (Zoph et al., 2016), adapting to a new language is conceptually simpler; given an NMT model pre-trained on a high-resource language pair (*parent*), we can just continue the training with bilingual data of another language pair (*child*). Here, the vocabulary mismatch between languages is still a problem, which seriously limits the performance especially for distant languages.

This work proposes three novel ideas to make transfer learning for NMT widely applicable to various languages:

- We alleviate the vocabulary mismatch between parent and child languages via cross-lingual word embedding.
- We train a more general encoder in the parent training by injecting artificial noises, making it easier for the child model to adapt to.
- We generate synthetic data from parallel data of the parent language pair, improving the low-resource transfer where the conventional back-translation (Sennrich et al., 2016b) fails.

These techniques give incremental improvements while we keep the transfer unsupervised, i.e. it does not require bilingual information between the transferor and the transferee. Note that adapting to a new language is done without shared vocabularies; we need neither to rearrange joint subword units nor to restart the parent model training.

Experiments show that our methods offer significant gain in translation performance up to +5.1% BLEU over plain transfer learning, even when transferring to an unrelated, low-resource

language. The results significantly outperform multilingual joint training (Johnson et al., 2017) in all of our experiments. We also provide in-depth analyses of the following aspects to understand the behavior of NMT transfer and maximize its performance: type of the pre-trained embedding, synthetic data generation methods, size of the transferred vocabulary, and parameter freezing.

## 2 Neural Machine Translation

Before describing our transfer learning approach, this section covers basics of an NMT model. Explanations here are not based on a specific architecture but extendable to more complex model variants.

For a source sentence $f_1^J = f_1, ..., f_j, ..., f_J$ (length $J$) and a corresponding target sentence $e_1^I = e_1, ..., e_i, ..., e_I$ (length $I$), NMT models the probability $p(e_1^I|f_1^J)$ with several components: source/target word embeddings, an encoder, a decoder, and an output layer.

Source word embedding $E^{\text{src}}$ maps a discrete word $f$ (as a one-hot vector) to a continuous representation (*embedding*) of that word $E^{\text{src}}(f)$. In practice, it is implemented by a lookup table and stored in a matrix in $\mathbb{R}^{D \times V^{\text{src}}}$, where $D$ is the number of dimensions of the embedding. Target word embedding is analogous.

An encoder takes a sequence of source word embeddings $E^{\text{src}}(f_1^J)$ and produces a sequence of hidden representations $\mathbf{h}_1^J$ for the source sentence. The encoder can be modeled with recurrent (Sutskever et al., 2014), convolutional (Gehring et al., 2017), or self-attentive layers (Vaswani et al., 2017). The encoder is responsible for modeling syntactic and semantic relationships among the source words, including word order.

A decoder generates target words for each target position $i$ from its internal state $\mathbf{s}_i$, which depends on $\mathbf{h}_1^J$, $E^{\text{tgt}}(e_{i-1})$, and $\mathbf{s}_{i-1}$. It keeps track of the generated hypothesis up to position $i$-1 and relates the generation with source representations $\mathbf{h}_1^J$. For shared vocabularies between source and target languages, the target embedding weights can be tied with the source embedding weights, i.e. $E^{\text{src}} = E^{\text{tgt}}$.

The model is trained on a parallel corpus by optimizing for the cross-entropy loss with the stochastic gradient descent algorithm. Translation is carried out with a beam search. For more details, we refer the reader to Bahdanau et al. (2015) and Vaswani et al. (2017).

## 3 Transfer Learning for NMT

In general, transfer learning is reusing the knowledge from other domains/tasks when facing a new problem (Thrun and Pratt, 2012). It has been of continued interest in machine learning for the past decades, especially when there is not enough training data for the problem at hand. Much attention is given to transfer learning for neural networks, since hidden layers of the network can implicitly learn general representations of data; the knowledge can be readily transferred by copying the hidden layer weights to another network (Caruana, 1995; Bengio, 2012).

For NMT, the easiest case of transfer learning is across text domains. Having an NMT model trained on some data, we can continue the training from the same network parameters with data from another domain (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016). Transfer from another natural language processing task is also straightforward; for example, we can initialize the parameters of NMT models with pre-trained language models of corresponding languages, since the encoder and decoder are essentially language models except a few additional translation-specific components (Ramachandran et al., 2017; Lample and Conneau, 2019).

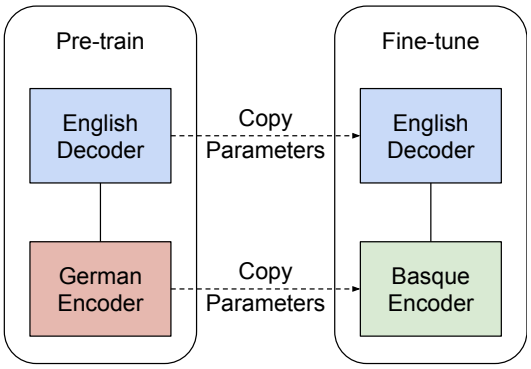


Figure 1: Diagram of transfer learning for NMT from German→English to Basque→English.

However, it is inherently difficult to transfer NMT models between languages, i.e. pre-train a model for a high-resource language pair and use the trained parameters for a low-resource language pair (Figure 1). Changing a language introduces a completely different data space that does not fit to the pre-trained model. In the following, we describe this discrepancy in detail and propose

our solutions. We focus on switching source languages, while the target language is fixed.

### 3.1 Cross-lingual Word Embedding

The biggest challenge of cross-lingual transfer is the vocabulary mismatch. A natural language vocabulary is discrete and unique for each language, while the mapping between two different vocabularies is non-deterministic and arbitrary. Therefore, when we merely replace a source language, the NMT encoder will see totally different input sequences; pre-trained encoder weights do not get along with the source embedding anymore.

A popular solution to this is sharing the vocabulary among the languages of concern (Nguyen and Chiang, 2017; Kocmi and Bojar, 2018). This is often implemented with joint learning of subword units (Sennrich et al., 2016c). Despite its effectiveness, it has an intrinsic problem in practice: A parent model must be trained already with a shared vocabulary with child languages. Such a pre-trained parent model can be transferred only to those child languages using the same shared vocabulary. When we adapt to a new language whose words are not included in the shared vocabulary, we should learn a joint subword space again with the new language and retrain the parent model accordingly—very inefficient and not scalable.

A shared vocabulary is also problematic in that it must be divided into language-specific portions. When many languages share it, an allocated portion for each will be smaller and accordingly less expressive. This is the reason why the vocabulary is usually shared only for linguistically related languages, effectively increasing the portion of common surface forms.

In this work, we propose to keep the vocabularies separate, but share their embedding spaces instead of surface forms. This can be done independently from the parent model training and requires only monolingual data of the child language:

1. Learn monolingual embedding of the child language $E^{\text{mono}}_{\text{child}}$, using e.g. the skip-gram algorithm (Mikolov et al., 2013).
2. Extract source embedding $E^{\text{src}}_{\text{parent}}$ from a pre-trained parent NMT model.
3. Learn a cross-lingual linear mapping $W \in \mathbb{R}^{D \times D}$ between 1 and 2 by minimizing the objective below:

$$\sum_{(f,f') \in S} \|W E^{\text{mono}}_{\text{child}}(f) - E^{\text{src}}_{\text{parent}}(f')\|_2 \quad (1)$$

4. Replace source embedding of the parent model parameters with the learned cross-lingual embedding.

$$E^{\text{src}}_{\text{parent}} \leftarrow W E^{\text{mono}}_{\text{child}} \quad (2)$$

5. Initialize the child model with 4 and start the NMT training on the child language pair.

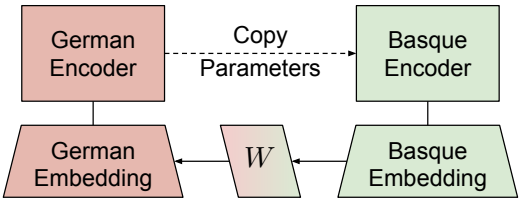


Figure 2: Cross-lingual mapping of a child (Basque) embedding to the parent (German) embedding.

The dictionary $S$ in Step 3 can be obtained in an unsupervised way by adversarial training (Conneau et al., 2018) or matching digits between the parent and child languages (Artetxe et al., 2017). The mapping $W$ can be also iteratively refined with self-induced dictionaries of mutual parent-child nearest neighbors (Artetxe et al., 2017), which is still unsupervised. The cross-lingually mapped child embeddings fit better as input to the parent encoder, since they are adjusted to a space similar to that of the parent input embeddings (Figure 2).

Note that in Step 4, the mapping $W$ is not explicitly inserted as additional parameters in the network. It is multiplied by $E^{\text{mono}}_{\text{child}}$ and the result is used as the initial source embedding weights. The initialized source embedding is also fine-tuned along with the other parameters in the last step.

These steps do not involve rearranging a joint vocabulary or retraining of the parent model. Using our method, one can pre-train a single parent model once and transfer it to many different child languages efficiently.

Our method is also effective for non-related languages that do not share surface forms, since we address the vocabulary mismatch in the embedding level. After each word is converted to its embedding, it is just a continuous-valued vector in a mathematical space; matching vocabularies is done by transforming the vectors irrespective of language-specific alphabets.

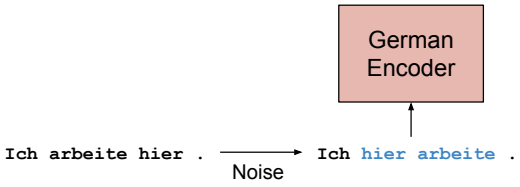


Figure 3: Injecting noise into a German (parent) source sentence.

### 3.2 Artificial Noises

Another main difference between languages is the word order, namely syntactic structure of sentences. Neural sequence-to-sequence models are highly dependent on sequential ordering of the input, i.e. absolute/relative positions of input tokens.

When we train an encoder for a language, it learns the language-specific word order conventions, e.g. position of a verb in a clause, structure of an adverb phrase, etc. If the input language is changed, the encoder should adjust itself to unfamiliar word orders. The adaptation gets more difficult for non-related languages.

To mitigate this syntactic difference in cross-lingual transfer for NMT, we suggest to generalize the parent encoder so that it is not overoptimized to the parent source language. We achieve this by modifying the source side of the parent training data, artificially changing its word orders with random noises (Figure 3). The noise function includes (Hill et al., 2016; Kim et al., 2018):

- Inserting a word between original words uniformly with a probability $p_{\text{ins}}$ at each position, choosing the inserted word uniformly from the top $V_{\text{ins}}$ frequent words
- Deleting original words uniformly with a probability $p_{\text{del}}$ at each position
- Permuting original word positions uniformly within a limited distance $d_{\text{per}}$

The noises are injected into every source sentence differently for each epoch. The encoder then sees not only word orders of the parent source language but also other various sentence structures. Since we set limits to the randomness of the noises, the encoder is still able to learn general monotonicity of natural language sentences. This makes it easier for the parent encoder to adapt to a child source language, effectively transferring the pre-trained language-agnostic knowledge of input sequence modeling.

### 3.3 Synthetic Data from Parent Model Training Data

Transfer learning for NMT is particularly necessary for low-resource language pairs where the bilingual data is scarce. The standard technique to address the scarcity is generating synthetic parallel data from target monolingual corpora via back-translation (Sennrich et al., 2016b). However, this works only if the generated source sentences are of sufficiently acceptable quality. In low-resource translation tasks, it is hard to train a good target-to-source translation model, which is used to produce the source hypotheses.

For these scenarios, we devise a simple trick to create additional parallel data for the child language pair without training a target-to-source translation model. The idea is to reuse the parallel data already used for training the parent model. In the source side, we retain only those tokens that exist in the child vocabulary and replace all other tokens with a predefined token, e.g. <unk> (Figure 4). The target side stays the same as we do not switch the languages.

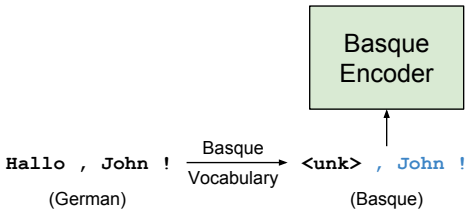


Figure 4: Synthetic Basque sentence generated from a German sentence.

The source side of this synthetic data consists only of the overlapping vocabulary entries between the parent and child languages. By including this data in the child model training, we prevent an abrupt change of the input to the pre-trained model while keeping the parent and child vocabularies separated. It also helps to avoid overfitting to a tiny parallel data of the child language pair.

In addition, we can expect a synergy with cross-lingual word embedding (Section 3.1), where the source embedding space of the child task is transformed into that of the parent task. In this cross-lingual space, an overlapping token between parent and child vocabularies should have a very similar embedding to that in the original parent embedding space, to which the pre-trained encoder is already familiar. This helps to realize a smooth

| Family | Source Language | Data (→English) [#sents] |
|---|---|---|
| Germanic | German | 10,111,758 |
| Isolate | Basque | 5,605 |
| Slavic | Slovenian | 17,103 |
| | Belarusian | 4,509 |
| Turkic | Azerbaijani | 5,946 |
| | Turkish | 9,998 |

Table 1: Language families and parallel data statistics.

transition from parent source input to child source input in the transfer process.

## 4 Main Results

We verify the effect of our techniques in transfer learning setups with five different child source languages: Basque (eu), Slovenian (sl), Belarusian (be), Azerbaijani (az), and Turkish (tr). Target language is fixed to English (en) and we use German→English as the parent language pair.

**Data**: The parent model was trained on parallel data of WMT 2018 news translation task[1] and synthetic data released by Sennrich et al. (2016a). For the child language pairs, we used IWSLT 2018 low-resource MT task data (eu-en) (Jan et al., 2018), IWSLT 2014 MT task data (sl-en) (Cettolo et al., 2014), TED talk data from (Qi et al., 2018) (be-en/az-en), and subsampling of WMT 2018 news translation task data (tr-en). Statistics of the parallel corpora are given in Table 1. Note that the child source languages are linguistically far from the parent source.

Every training dataset was preprocessed with the Moses tokenizer[2], where the source side was lowercased and the target side was frequent-cased.

**Transfer learning**: All NMT models in our experiments follow the base 6-layer Transformer architecture of Vaswani et al. (2017), except that the source and target embedding weights are not tied. Each source language was encoded with byte pair encoding (BPE) (Sennrich et al., 2016c) with 20k merge operations, while the target language was encoded with 50k BPE merges. Dropout with probability of 0.3 was applied to Transformer pre-post/activation/attention components in both parent and child model trainings. Training was carried out with Sockeye (Hieber et al., 2017) using the Adam optimizer (Kingma and Ba, 2014) with the default parameters. The maximum sentence length was set to 100 and the batch size to 4,096 words. We stopped the training when perplexity on a validation set was not improving for 12 checkpoints. We set checkpoint frequency to 10,000 updates for the parent model and 1,000 updates for the child models. The parent model yields 39.2% BLEU on WMT German→English newstest2016 test set.

**Baseline**: As a baseline child model without transfer learning, we used the same setting as above but learned a shared source-target BPE vocabulary with 20k merge operations. We also tied source and target embeddings as suggested for low-resource settings in Schamper et al. (2018). Dropout was applied also to the embedding weights for the baselines.

**Multilingual**: We also compare our transfer learning with the multilingual training where a single, shared NMT model is trained for the parent and child language pairs together from scratch (Johnson et al., 2017). For each child task, we learned a joint BPE vocabulary of all source and target languages in the parent/child tasks with 32k merge operations. The training data for the child task was oversampled so that each mini-batch has roughly 1:1 ratio of the parent/child training examples.

Note that we built a different multilingual model for each child task. Since they depend on shared vocabularies, we should restructure the vocabulary and retrain the model for each of the new language pairs we wish to adapt to.

**Cross-lingual word embedding**: To pre-train word embeddings, we used Wikimedia dumps[3] of timestamp 2018-11-01 for all child languages except Turkish for which we used WMT News Crawl 2016-2017. From Wikimedia dumps, the actual articles were extracted first[4], which were split to sentences using the StanfordCoreNLP toolkit (Manning et al., 2014). Monolingual embeddings were trained with fasttext (Bojanowski et al., 2017) with minimum word count 0. For learning the cross-lingual mappings, we ran 10 epochs of adversarial training and another 10 epochs of dictionary-based refinement using MUSE (Con-

[1] http://www.statmt.org/wmt18/translation-task.html
[2] http://www.statmt.org/moses/
[3] https://dumps.wikimedia.org/
[4] https://github.com/attardi/wikiextractor/

| System | BLEU [%] eu-en | sl-en | be-en | az-en | tr-en |
|---|---|---|---|---|---|
| Baseline | 1.7 | 10.1 | 3.2 | 3.1 | 0.8 |
| Multilingual (Johnson et al., 2017) | 5.1 | 16.7 | 4.2 | 4.5 | 8.7 |
| Transfer (Zoph et al., 2016) | 4.9 | 19.2 | 8.9 | 5.3 | 7.4 |
| + Cross-lingual word embedding | 7.4 | 20.6 | 12.2 | 7.4 | 9.4 |
| + Artificial noises | 8.2 | 21.3 | 12.8 | 8.1 | 10.1 |
| + Synthetic data | **9.7** | **22.1** | **14.0** | **9.0** | **11.3** |

Table 2: Translation results of different transfer learning setups.

neau et al., 2018). We chose top 20k types as discriminator inputs and 10k as maximum dictionary rank.

**Artificial noises**: Following Kim et al. (2018), we used these values for the noise model: $p_{\text{ins}} = 0.1$, $V_{ins} = 50$, $p_{\text{del}} = 0.1$, and $d_{\text{per}} = 3$. We empirically found that these values are optimal also for our purpose. The parent model trained with noises gives 38.2% BLEU in WMT German→English newstest2016: 1.0% worse than without noises.

**Synthetic data**: We uniformly sampled 1M sentence pairs from German→English parallel data used for the parent training and processed them according to Section 3.3. The child model parallel data was oversampled to 500k sentence pairs, making an overall ratio of 1:2 between the parallel and synthetic data. We also tried other ratio values, e.g. 1:1, 1:4, or 2:1, but the performance was consistently worse.

Table 2 presents the results. Plain transfer learning already gives a boost but is still far from a satisfying quality, especially for Basque→-English and Azerbaijani→English. On top of that, each of our three techniques offers clear, incremental improvements in all child language pairs with a maximum of 5.1% BLEU in total.

Cross-lingual word embedding shows a huge improvement up to +3.3% BLEU, which exhibits the strength of connecting parent-child vocabularies on the embedding level. If we train the parent model with artificial noises on the source side, the performance is consistently increased by up to +0.8% BLEU. This occurs even when dropout is used in the parent model training; randomizing word orders provides meaningful regularization which cannot be achieved via dropout. Finally, our synthetic data extracted from the parent parallel data is proved to be effective in low-resource transfer to substantially different languages: We obtain an additional gain of at most +1.5% BLEU.

Our results also surpass the multilingual joint training by a large margin in all tasks. One shared model for multiple language pairs inherently limits the modeling capacity for each task. Particularly, if one language pair has much smaller training data than the other, oversampling the low-resource portion is not enough to compensate the scale discrepancy in multilingual training. Transfer learning with our add-on techniques is more efficient to exploit knowledge of high-resource language pairs and fine-tune the performance towards a child task.

## 5 Analysis

In this section, we further investigate our methods in detail in comparison to their similar variants, and also perform ablation studies for the NMT transfer in general.

### 5.1 Types of Pre-trained Embedding

| Pre-trained embedding | BLEU [%] |
|---|---|
| None | 5.3 |
| Monolingual | 6.3 |
| Cross-lingual (az-de) | **7.4** |
| Cross-lingual (az-en) | 7.1 |

Table 3: Azerbaijani→English translation results with different types of pre-trained source embeddings.

We analyze the effect of the cross-linguality of pre-trained embeddings in Table 3. We observe that monolingual embedding without a cross-lingual mapping also improves the transfer learning, but is significantly worse than our proposed embedding, i.e. mapped to the parent source (de)

embedding. The mapping can be learned also with the target (en) side with the same procedure as in Section 3.1. The target-mapped embedding is not compatible with the pre-trained encoder but directly guides the child model to establish the connection between the new source and the target. It also improves the system, but our method is still the best among the three embedding types.

### 5.2 Synthetic Data Generation

| Synthetic data | BLEU [%] |
|---|---|
| None | 8.2 |
| Back-translation | 8.3 |
| Empty source | 8.2 |
| Copied target | 8.9 |
| Parent model data | **9.7** |
| + Cross-lingual replacement | 8.7 |

Table 4: Basque→English translation results with synthetic data generated using different methods.

In Table 4, we compare our technique in Section 3.3 with other methods of generating synthetic data. For a fair comparison, we used the same target side corpus (1M sentences) for all these methods.

As explained in Section 3.3, back-translation (Sennrich et al., 2016b) is not beneficial here because the generated source is of too low quality. Empty source sentence is proposed along with back-translation as its simplification, which does not help either in transfer learning. Copying target sentences to the source side is yet another easy way to obtain synthetic data (Currey et al., 2017). It gives an improvement to a certain extent; however, our method of using the parent model data works much better in transfer learning.

We manually looked at the survived tokens in the source side of our synthetic data. We observed lots of overlapping tokens over the parent and child source vocabularies even if they were not shared: 4,487 vocabulary entries between Basque and German. Approximately 2% of them are punctuation symbols and special tokens, 7% are digits, and 62% are made of Latin alphabets, a large portion of which is devoted to English words (e.g. named entities) or their parts. The rest of the vocabulary is mostly of noisy tokens with exotic alphabets.

As Figure 4 illustrates, just punctuation symbols and named entities can already define a basic structure of the original source sentence. Such tokens play the role of anchors in translation; they are sure to be copied to the target side. The surrounding `<unk>` tokens are spread according to the source language structure, whereas merely copying the target sentence to the source (Currey et al., 2017) ignores the structural difference between source and target sentences. Note that our trick applies also to the languages with completely different alphabets, e.g. Belarusian and German (see Table 2).

We also tested an additional processing for our synthetic data to reduce the number of unknown tokens. We replaced non-overlapping tokens in the German source side with the closest Basque token in the cross-lingual word embedding space. The result is, however, worse than not replacing them; we noticed that this subword-by-subword translation produces many Basque phrases with wrong BPE merges (Kim et al., 2018).

### 5.3 Vocabulary Size

| | BLEU [%] | |
|---|---|---|
| BPE merges | sl-en | be-en |
| 10k | **21.0** | 11.2 |
| 20k | 20.6 | **12.2** |
| 50k | 20.2 | 10.9 |
| 70k | 20.0 | 10.9 |

Table 5: Translation results with different sizes of the source vocabulary.

Table 5 estimates how large the vocabulary should be for the language-switching side in NMT transfer. We varied the number of BPE merges on the source side, fixing the target vocabulary to 50k merges. The best results are with 10k or 20k of BPE merges, which shows that the source vocabulary should be reasonably small to maximize the transfer performance. Less BPE merges lead to more language-independent tokens; it is easier for the cross-lingual embedding to find the overlaps in the shared semantic space.

If the vocabulary is excessively small, we might lose too much language-specific details that are necessary for the translation process. This is shown in the 10k merges of Belarusian→English.

### 5.4 Freezing Parameters

Lastly, we conducted an ablation study of freezing parent model parameters in the child training

| Frozen parameters | BLEU [%] |
|---|---|
| None | 21.0 |
| Target embedding | 21.4 |
| + Target self-attention | **22.1** |
| + Encoder-decoder attention | 21.8 |
| + Feedforward sublayer | 21.3 |
| + Output layer | 21.9 |

Table 6: Slovenian→English translation results with freezing different components of the decoder.

process (Table 6). We show only the results when freezing the decoder; in our experiments, freezing any component of the encoder always degrades the translation performance. The experiments were done at the final stage with all of our three proposed methods applied.

Target embedding and target self-attention parts are independent of the source information, so it makes sense to freeze those parameters even when the source language is changed. On the contrary, encoder-decoder attention represents the relation between source and target sentences, so it should be redefined for a new source language. The performance deteriorates when freezing feedforward sublayers, since it is directly influenced by the encoder-decoder attention layer. The last row means that we freeze all parameters of the decoder; it is actually better than freezing all but the output layer.

## 6 Related Work

Transfer learning is first introduced for NMT in Zoph et al. (2016), yet with a small RNN architecture and on top frequent words instead of using subword units. Nguyen and Chiang (2017) and Kocmi and Bojar (2018) use shared vocabularies of BPE tokens to improve the transfer learning, but this requires retraining of the parent model whenever we transfer to a new child language.

Multilingual NMT trains a single model with parallel data of various translation directions jointly from scratch (Dong et al., 2015; Johnson et al., 2017; Firat et al., 2016; Gu et al., 2018). Their methods also rely on shared subword vocabularies so it is hard for their model to adapt to a new language.

Cross-lingual word embedding is studied for the usages in MT as follows. In phrase-based SMT, Alkhouli et al. (2014) builds translation models with word/phrase embeddings. Kim et al. (2018) uses cross-lingual word embedding as a basic translation model for unsupervised MT and attach other components on top of it. Artetxe et al. (2018c) and Lample et al. (2018a) initialize their unsupervised NMT models with pre-trained cross-lingual word embeddings. Qi et al. (2018) do the same initialization for supervised cases, observing only improvements in multilingual setups.

Artificial noises for the source sentences are used to counteract word-by-word training data in unsupervised MT (Artetxe et al., 2018c; Lample et al., 2018a; Kim et al., 2018), but in this work, they are used to regularize the NMT.

Neubig and Hu (2018) study adapting a multilingual NMT system to a new language. They train for a child language pair with additional parallel data of its similar language pair. Our synthetic data method does not rely on the relatedness of languages but still shows a good performance. They learn just a separate subword vocabulary for the child language without a further care, which we counteract with cross-lingual word embedding.

Sachan and Neubig (2018) show ablation studies on parameter sharing and freezing in one-to-many multilingual setup with shared vocabularies. Our work conduct the similar experiments in the transfer learning setting with separate vocabularies.

Platanios et al. (2018) augment a multilingual model with language-specific embeddings from which the encoder and decoder parameters are inferred with additional linear transformations. They only mention its potential to transfer to an unseen language without any results on it. Our work focuses on transferring a pre-trained model to a new language without any change in the model architecture but with an explicit guidance for cross-linguality on the word embedding level.

Wang et al. (2019) address the vocabulary mismatch in multilingual NMT by using shared embeddings of character $n$-grams and common semantic concepts. Their method has a strict assumption that the languages should be related orthographically with shared alphabets, while our method is not limited to similar languages and directly benefits from advances in cross-lingual word embedding for distant languages.

Another line of research on low-resource MT is unsupervised learning (Lample et al., 2018a,b; Lample and Conneau, 2019; Artetxe et al.,

2018b,c; Kim et al., 2018), training translation models only with monolingual data. However, these methods are verified mostly in high-resource language pairs, e.g. French↔English, where there is no need to restrict the training data to only monolingual corpora. In low-resource language pairs with little linguistic similarity, Neubig and Hu (2018) and Guzmán et al. (2019) show that unsupervised MT methods do not function at all. We tested an unsupervised MT software Lample and Conneau (2019) internally, which also resulted in failure, e.g. 1% BLEU at the Basque→English task of Section 4. Moreover, unsupervised MT methods usually require a very long training time—at least 1-2 weeks with a single GPU—due to its iterative nature, while our cross-lingual transfer needs only a couple of hours of training once you have a parent model.

Alternatively, one might consider using parallel data involving a pivot language, either by decoding in two consecutive steps (Kauers et al., 2002; De Gispert and Marino, 2006; Utiyama and Isahara, 2007; Costa-Jussà et al., 2011) or by creating pivot-based synthetic data (De Gispert and Marino, 2006; Bertoldi et al., 2008; Zheng et al., 2017; Chen et al., 2017). These methods cannot be applied to most of the language pairs from/to English, because it is extremely difficult to collect parallel data with another third language other than English.

## 7 Conclusion

In this paper, we address the problem of transferring an NMT model to unseen, unrelated language pairs. We propose three novel techniques to improve the transfer without vocabulary sharing between parent and child source languages.

Firstly, we transform monolingual embeddings of the new language into the embedding space of the parent NMT model. This accomplishes an effective transition of vocabularies on the embedding level. Secondly, we randomize the word orders in the parent model training to avoid overfitting to the parent source language. This makes it easier for the encoder to adapt to the new language syntax. For the first time, we show a practical usage of artificial noises to regularize an NMT model. Lastly, we reuse parallel data of the parent language pair in the child training phase to avoid an abrupt change of the training data distribution.

All three methods significantly improve over plain transfer learning with a total gain of up to +5.1% BLEU in our experiments, consistently outperforming multilingual joint training. Our methods do not require retraining of a shared vocabulary or the parent model, enabling an incremental transfer of the same parent model to various (possibly unrelated) languages. Our implementation of the proposed methods is available online.[5]

As for future work, we will test our methods in the NMT transfer where the target language is switched. We also plan to compare different algorithms for learning the cross-lingual mapping (Artetxe et al., 2018a; Xu et al., 2018; Joulin et al., 2018) to optimize the transfer performance.

## Acknowledgments

This work has received funding from the European Research Council (ERC) (under the European Union's Horizon 2020 research and innovation programme, grant agreement No 694537, project "SEQCLAS") and the Deutsche Forschungsgemeinschaft (DFG; grant agreement NE 572/8-1, project "CoreTec"). The GPU cluster used for the experiments was partially funded by DFG Grant INST 222/1168-1. The work reflects only the authors' views and none of the funding agencies is responsible for any use that may be made of the information it contains.

## References

Tamer Alkhouli, Andreas Guta, and Hermann Ney. 2014. Vector space models for phrase-based machine translation. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 1–10.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, volume 1, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 789–798.

[5] https://github.com/yunsukim86/sockeye-transfer

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

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. In *Proceedings of 6th International Conference on Learning Representations (ICLR 2018)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 2015 International Conference on Learning Representations (ICLR 2015)*.

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. In *International Workshop on Spoken Language Translation (IWSLT) 2008*.

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.

Rich Caruana. 1995. Learning many related tasks at the same time with backpropagation. In *Advances in neural information processing systems*, pages 657–664.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT 2011)*, pages 2–17, Hanoi, Vietnam.

Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of 6th International Conference on Learning Representations (ICLR 2018)*.

Marta R Costa-Jussà, Carlos Henríquez, and Rafael E Banchs. 2011. Enhancing scarce-resource language translation through pivot combinations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1361–1365.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Adrià De Gispert and Jose B Marino. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1723–1732.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 866–875.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv:1612.06897*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 1243–1252.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 344–354.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv:1902.01382*.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv:1712.05690*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 1367–1377.

Niehues Jan, Roldano Cattoni, Stüker Sebastian, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The iwslt 2018 evaluation campaign. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, pages 2–6.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics (TACL)*, 5(1):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.

Manuel Kauers, Stephan Vogel, Christian Fügen, and Alex Waibel. 2002. Interlingua based statistical machine translation. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*.

Yunsu Kim, Jiahui Geng, and Hermann Ney. 2018. Improving unsupervised word-by-word translation with language model and denoising autoencoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 862–868. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the 1st ACL Workshop on Neural Machine Translation (WNMT 2017)*, pages 28–39.

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

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of 6th International Conference on Learning Representations (ICLR 2018)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, pages 76–79.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

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.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 296–301.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435.

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)*, volume 2, pages 529–535.

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271. Association for Computational Linguistics.

Julian Schamper, Jan Rosendahl, Parnia Bahar, Yunsu Kim, Arne Nix, and Hermann Ney. 2018. The rwth aachen university supervised machine translation systems for wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 496–503.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, volume 1, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 3104–3112. MIT Press.

Sebastian Thrun and Lorien Pratt. 2012. *Learning to learn*. Springer Science & Business Media.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *Proceedings of the 2019 International Conference on Learning Representations (ICLR 2019)*.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474.

Hao Zheng, Yong Cheng, and Yang Liu. 2017. Maximum expected likelihood estimation for zero-resource neural machine translation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4251–4257. AAAI Press.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 1568–1575.