In Neural Machine Translation, What Does Transfer Learning Transfer?

Citation for published version:
Aji, AF, Bogoychev, N, Heafield, K & Sennrich, R 2020, In Neural Machine Translation, What Does Transfer Learning Transfer? in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics (ACL), pp. 7701–7710, 2020 Annual Conference of the Association for Computational Linguistics, Virtual conference, United States, 5/07/20. https://doi.org/10.18653/v1/2020.acl-main.688

Digital Object Identifier (DOI):
10.18653/v1/2020.acl-main.688

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

In Neural Machine Translation, What Does Transfer Learning Transfer?

Alham Fikri Aji[1] **Nikolay Bogoychev**[1] **Kenneth Heafield**[1] **Rico Sennrich**[2,1]

[1]School of Informatics, University of Edinburgh
[2]Department of Computational Linguistics, University of Zurich
{a.fikri, n.bogoych, kenneth.heafield, rico.sennrich}@ed.ac.uk

Abstract

Transfer learning improves quality for low-resource machine translation, but it is unclear what exactly it transfers. We perform several ablation studies that limit information transfer, then measure the quality impact across three language pairs to gain a black-box understanding of transfer learning. Word embeddings play an important role in transfer learning, particularly if they are properly aligned. Although transfer learning can be performed without embeddings, results are sub-optimal. In contrast, transferring only the embeddings but nothing else yields catastrophic results. We then investigate diagonal alignments with auto-encoders over real languages and randomly generated sequences, finding even randomly generated sequences as parents yield noticeable but smaller gains. Finally, transfer learning can eliminate the need for a warm-up phase when training transformer models in high resource language pairs.

1 Introduction

Transfer learning is a common method for low-resource neural machine translation (NMT) (Zoph et al., 2016; Dabre et al., 2017; Qi et al., 2018; Nguyen and Chiang, 2017; Gu et al., 2018b). However, it is unclear what settings make transfer learning successful and what knowledge is being transferred.

Understanding why transfer learning is successful can improve best practices while also opening the door to investigating ways to gain similar benefits without requiring parent models. In this paper, we perform several ablation studies on transfer learning in order to understand what information is being transferred.

We apply a black box methodology by measuring the quality of end-to-end translation systems. Typically, our experiments have a baseline that was trained from scratch, an off-the-shelf transfer learning baseline and simplified versions of the transfer learning scheme. If a simplified version recovers some of the quality gains of full transfer learning, it suggests that the simplified version has captured some of the information being transferred. Since information may be transferred redundantly, our claims are limited to sufficiency rather than exclusivity.

Transferring word embeddings is not straightforward since languages have different vocabularies. Zoph et al. (2016) claimed that vocabulary alignment is not necessary, while Nguyen and Chiang (2017) and Kocmi and Bojar (2018) suggest a joint vocabulary. We find that the vocabulary has to be aligned before transferring the embedding to achieve a substantial improvement. Transfer learning without the embedding or with vocabulary mismatches is still possible, but with lower quality. Conversely, transferring only the word embeddings can be worse than transferring nothing at all.

A rudimentary model of machine translation consists of alignment and token mapping. We hypothesize that these capabilities are transferred across languages. To test this, we experiment with transferring from auto-encoders that learn purely diagonal alignment and possibly language modelling. To remove the effect of language modelling, we train auto-encoders on random strings sampled uniformly. However, all of these scenarios still have simple copying behaviour, especially with tied embeddings. Therefore, we also attempt a bijective vocabulary mapping from source to target, forcing the model to learn the mapping as well. Curiously, parents trained with bijectively-mapped vocabularies transfer slightly better to children.

We then investigate transfer learning for high-resource children, where the goal is reduced training time since they mainly attain the same quality. Transfer learning primarily replaces the warm-up period, though only real language parents yielded faster training.

2 Related Work

Transfer learning has been successfully used in low-resource scenarios for NMT. Zoph et al. (2016) gain 5 BLEU points in Uzbek–English by transferring from French–English. Their style of transfer learning copies the entire model, including word embeddings, ignoring the vocabulary mismatch between parent and child. They used separate embeddings for source and target language words, whereas tied embeddings (Press and Wolf, 2017; Vaswani et al., 2017) have since become the de-facto standard in low-resource NMT. Tied embeddings provide us with the opportunity to revisit some of their findings. In Section 5, we find an English–English copy model does work as a parent with tied embeddings, whereas Zoph et al. (2016) reported no gains from a copy model with untied embeddings.

Methods to cope with vocabulary mismatch have improved since Zoph et al. (2016). Kocmi and Bojar (2018) suggest that a shared vocabulary between the parent language and the child is beneficial, though this requires knowledge of the child languages when the parent is trained. Addressing this issue, Gheini and May (2019) proposed a universal vocabulary for transfer learning. Their universal vocabulary was obtained by jointly training the sub-word tokens across multiple languages at once, applying Romanisation to languages in non-Latin scripts. However, unseen languages may only be representable in this universal vocabulary with a very aggressive and potentially sub-optimal subword segmentation. Orthogonally, Kim et al. (2018); Lample et al. (2018); Artetxe et al. (2018); Kim et al. (2019) use bilingual word embedding alignment to initialise the embedding layer to tackle low resource language pairs. In Section 4.2, we compare a variety of vocabulary transfer methods.

Prior work (Dabre et al., 2017; Nguyen and Chiang, 2017) stated that a related language is the best parent for transfer learning. Lin et al. (2019) explore options to choose the best parent and conclude that the best parent language might not necessarily be related but is instead based on external factors such as the corpus size. In Section 3, we try two parent models in both directions to set baselines for the rest of the paper; an exhaustive search is not our main purpose.

Another approach to low-resource (or even zero-shot) NMT is through multilingual models (Johnson et al., 2016), which is similar to training the parent and child simultaneously. A related idea creates meta-models with vocabulary residing in a shared semantic space (Gu et al., 2018a,b).

If there is more parallel data with a third language, often English, then pivoting through a third language can outperform direct translation (Cheng et al., 2016). This approach requires enough source–pivot and target–pivot parallel data, which is arguably hard in many low resource scenarios, such as Burmese, Indonesian, and Turkish.

Orthogonal to transfer learning, Lample et al. (2018) and Artetxe et al. (2018) have proposed a fully zero-shot approach for low resource languages that relies on aligning separately-trained word embeddings to induce an initial bilingual dictionary. The dictionary is then used as the basis for a translation model. However, these methods do not generalise to arbitrary language pairs (Søgaard et al., 2018). Moreover, our setting presumes a small amount of parallel data in the low-resource pair.

3 Baseline Transfer Learning

We start with arguably the simplest form of transfer learning: train a parent model then switch to training with the child's dataset following Zoph et al. (2016). We attempt to initialise the embedding vectors of the same tokens from the parent to the child. We later investigate different approaches to transferring the embeddings. As transfer learning requires a parent model, we start by sweeping different high-resource languages for the parent model to set a baseline.

Choosing a parent language pair is one of the first issues to solve when performing a transfer-learning experiment. However, this is not a simple task. Prior work (Dabre et al., 2017; Nguyen and Chiang, 2017) suggest that a related language is the best option, albeit related is not necessarily well defined. Recently, Lin et al. (2019) performed a grid-search across various parent languages to determine the best criteria for selecting the optimal parent when performing transfer learning. Their work showed that the best language parents might also be determined by external factors such as the corpus size, on top of the language relatedness. According to the BLEU score, the difference between various parents is usually not that significant.

We first explore four potential parents: German and Russian from/to English. From each of them, we transfer the parameters to our low-resource language pair of {Burmese, Indonesian, Turkish} to English. Before presenting the results, we lay out the experimental setup used for the rest of the paper.

3.1 High-Resource Datasets

We use German-English and Russian-English datasets for our parent models. Our German-English dataset is taken from the WMT17 news translation task (Bojar et al., 2017). Our Russian-English is taken from the WMT18 task (Bojar et al., 2018). For both pairs, we preprocess the input with byte-pair encoding (Sennrich et al., 2016b).

3.2 Low-Resource Datasets

We use the following datasets:

Burmese–English: For our My→En parallel data, we used 18k parallel sentences from the Asian Language Treebank (ALT) Project (Ding et al., 2018, 2019) collected from news articles.

Indonesian–English: Id→En parallel data consists of 22k news-related sentences, which are taken from the PAN Localization BPPT corpus.[1] This dataset does not have a test/validation split. Hence we randomly sample 2000 sentences to use as test and validation sets. We augment our data by back-translating (Sennrich et al., 2016a) News Crawl from 2015. Our total training set (including the back-translated sentences) consists of 88k pairs of sentences.

Turkish–English: Tr→En data comes from the WMT17 news translation task (Bojar et al., 2017). This data consists of 207k pairs of sentences. Similar to Id→En, we add a back-translation corpus from News Crawl 2015. Our total training data consists of 415k sentence pairs.

For all language pairs, we use byte-pair encoding (Sennrich et al., 2016b) to tokenise words into subword units.

3.3 Training Setup

We use a standard transformer-base architecture with six encoder and six decoder layers for all experiments with the default hyper-parameters (Vaswani et al., 2017). Training and decoding use Marian (Junczys-Dowmunt et al., 2018), while evaluation uses SacreBLEU (Post, 2018).

3.4 Results

	BLEU		
Parent	My→En	Id→En	Tr→En
-	4.0	20.6	19.0
En→De	17.5	27.5	20.2
En→Ru	17.8	27.4	20.3
De→En	17.3	26.3	20.1
Ru→En	17.1	26.8	20.6

Table 1: Transfer learning performance across different language parents.

Our results on Table 1 show that there is no clear evidence that one parent is better than another. Whether the non-English languages share a script or English is on the same side does not have a consistent impact. The main goal of this section was to set appropriate baselines; we primarily use English→German and German→English as the parents.

4 Transferring Embedding Information

Parent and child languages have a different vocabulary, so embeddings are not inherently transferable. We investigate what is transferred in the embeddings and evaluate several vocabulary combination methods.

4.1 Are the Embeddings Transferable?

We first explore whether the embedding matrix contains any transferable information. We divide the model into embedding parameters and everything else: inner layers. Table 2 shows what happens when these parts are or are not transferred.

Our low-resource languages achieve better BLEU even if we only transfer the inner layers. In contrast, only transferring the embeddings is not beneficial, and sometimes it is even harmful to the performance. Finally, transferring all layers yields the best performance.

To further investigate which part of the network is more crucial to transfer, we took the best-performing child then reset either the embeddings or inner layers and restarted training. We explore whether the model is capable of recovering the same or comparable quality by retraining. We can look at this experiment as ‘self’ transfer learning. Results are shown in Table 3. When the inner layers are reset, self-transfer performs poorly (close to the quality without transfer learning at all), even though the embeddings are properly transferred.

[1] http://www.panl10n.net/english/OutputsIndonesia2.htm

Transferring		BLEU						
		De→En parent			En→De parent			
Emb.	Inner	My→En	Id→En	Tr→En	My→En	Id→En	Tr→En	avg.
Y	Y	17.8	27.4	20.3	17.5	27.5	20.2	21.7
N	Y	13.6	25.3	19.4	10.8	24.9	19.3	18.3
Y	N	3.0	18.2	19.1	3.4	18.8	18.9	13.7
N	N	4.0	20.6	19.0	4.0	20.6	19.0	14.5

Table 2: Transfer learning performance by only transferring parts of the network. Inner layers are the non-embedding layers. N = not-transferred. Y = transferred.

	BLEU		
Transfer	My→En	Id→En	Tr→En
baseline (no transfer)	4.0	20.6	19.0
transfer, train	17.8	27.4	20.3
transfer, train, reset emb, train	13.3	25.0	20.0
transfer, train, reset inner, train	3.6	18.0	19.1

Table 3: Investigating the model's capability to restore its quality if we reset the parameters. We use En→De as the parent.

Conversely, the models can somewhat restore their quality even if we reset the embedding layer. This result further verifies that transferring the inner layers is the most critical aspect of transfer learning.

We conclude that transferring the inner layers is critical to performance, with far more impact than transferring the embeddings. However, the embedding matrix has transferable information, as long as the inner layers are included.

4.2 How to Transfer the Embeddings

Mixed recommendations exist on how to transfer embeddings between languages with different vocabularies. We compare methods from previous work, namely random assignment (Zoph et al., 2016) and joint vocabularies (Nguyen and Chiang, 2017) with two additional embedding assignment strategies based on the frequency and token matching as a comparison. In detail, we explore:

- **Exclude Embedding:** We do not transfer the embeddings at all. As such, we show that transfer learning works without transferring the embedding layer. In the present experiment, this method acts as one of the baselines.

- **Frequency Assignment:** We can transfer the embedding information regardless of the vocabulary mismatch. However, the toolkit sorts the words based on their frequency; therefore, embeddings are also transferred in that particular order. Regardless, we can determine whether word frequency information is transferred.

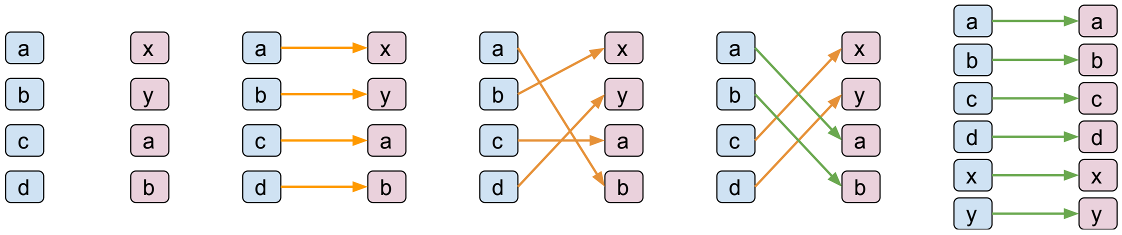
- **Random Assignment:** Zoph et al. (2016) suggest that randomly assigning a parent word embedding to each child word is sufficient, relying on the model to untangle the permutation. This approach is simple and language-agnostic, thus universally applicable. We shuffle the vocabulary to achieve a random assignment.

- **Joint Vocabulary:** Nguyen and Chiang (2017) suggest that it is better to use a shared vocabulary between the parent and child language. This can be obtained by training a joint BPE token. To achieve this, we transfer the word embedding information of the common tokens. Since tied embeddings are used, we share the same vocabulary between the target and source of both the parent and the child language. One drawback of this technique is that we must prepare the vocabulary in advance. Therefore, switching the parent or the child might require us to re-train the model.

- **Token Matching:** We assign the embeddings with the same token first and randomise the rest. This approach is designed to allow some word embeddings to be transferred correctly without the need to re-train the parent with every experiment, as in the case of joint vocabulary.

The different strategies are illustrated in Figure 1.

Prior experiments in Section 4.1 demonstrate that we can apply transfer learning even if we only transfer the inner layers. Curiously, random assignment and frequency assignment are not better than excluding the embeddings, except for Burmese to



(a) Exclude embedding (b) Freq. assignment (c) Random assignment (d) Token Match (e) Joint vocab

Figure 1: Illustration of various strategies on how to transfer the embedding vector.

	BLEU						
	De→En parent			En→De parent			
Embedding	My→En	Id→En	Tr→En	My→En	Id→En	Tr→En	avg.
-	4.0	20.6	19	4.0	20.6	19	14.5
Exclude embedding	13.6	25.3	19.4	10.8	24.9	19.3	18.3
Frequency assign	14.2	24.4	19.4	13.9	24.3	19.4	19.2
Random assign	13.9	24.6	19.2	13.8	23.9	19.3	19.0
Token matching	17.8	27.4	20.3	17.5	27.5	20.2	21.7
Joint vocabulary	18.5	27.5	20.9	18.5	28.0	19.6	22.0

Table 4: Transfer learning performance with different ways to handle the embedding layer.

English transferred from English to German. Therefore, the information in the embedding is lost when transferred to the incorrect token. From these results, we conclude that the model is incapable of untangling the embedding permutation as stated by Zoph et al. (2016).

Transfer learning yields better results when we attempt to transfer the embeddings to the correct tokens. In the joint vocabulary setting, not every token is observed in the parent language dataset; therefore, only a section of the embedding layer is correctly trained. However, we still observe a significant improvement over the random and frequency-based assignment.

We can also transfer the embedding vectors by matching and assigning the word embedding with the same tokens. Vocab matching achieves comparable results to joint vocabulary, except for the lowest-resource language, Burmese. Therefore, this simple matching can be used as a cheaper alternative over a joint vocabulary. On top of that, this approach is more efficient as we do not transfer and wastefully reserve extra memory for tokens that will not be seen in the child language.

These results suggest that word information stored in the embedding layer is transferable, as long as the vectors are assigned correctly. Therefore, better ways of handling the embedding layer transfer are joint BPE and token matching, as they further improve the performance of the child language pair.

5 Transferring Structural Information

To understand what information is being transferred with transfer learning, we test the parent model's performance on the child language without any additional training.

When a pre-trained model is transferred to another language pair, the model has not yet seen the child language vocabulary. When presented with an input in a new language, the model is unable to translate correctly. However, as we can see in Table 5, the model manages to perform diagonal alignment properly, albeit it is mostly copying the input (on average of 75% of the time).

Based on this observation, we see that fallback copying behaviour, including monotonic alignment, is transferred. This can be useful for named entity translation (Currey et al., 2017). To test our claim, we prepare parents that implicitly learn to copy or transform input tokens diagonally.

We can create a copy sequence model (or autoencoder) model by giving the model the same sentences for both source and target. We pick an English monolingual dataset. We also use a Chinese monolingual corpus to explore whether the chosen

Parent	Shared	Example
En→De	Id→En	src: Bank Mandiri bisa masuk dari mikro hingga korporasi . out: Bank Mandiri bisa memperingatkan dari cen@@ hingga korporasi . alignment: 0-0 1-1 3-3 5-5 6-6 7-7 9-2 9-4 9-8 9-9
De→En	Id→En	src: Bank Mandiri bisa masuk dari mikro hingga korporasi . out: seperti Mandiri bisa masuk a mikro hingga korporasi . alignment: 2-2 3-0 3-1 3-3 3-9 5-5 6-6 7-7 7-8 9-4

Table 5: Output example of transferred model without fine tuning. The model performs monotonic alignment.

monolingual language matters. Besides, we can artificially create a random sequence for the training set. The random sequence is useful to determine whether any language-specific information is being transferred, as such information is absent in a random sequence.

To simulate the translation behaviour better, we also prepare a substitution parallel corpus. We transform every token into another based on a predetermined 1:1 mapping. We create a substitution corpus for both the English and the synthetic corpus. With tied embeddings, the substitution corpus should help the model translate one token into another, instead of just copying. Table 6 illustrates the 6 monolingual/synthetic parents that we use for this experiment.

We perform transfer learning experiments from every monolingual and synthetic parent to all three child languages, as summarised in Table 7. For comparison, we also provide the result of transfer learning with an actual translation model as a parent. We notice that there is no improvement in transfer learning for the Turkish model in terms of the final BLEU. However, upon further investigation, transfer learning has an impact on the convergence speed, thus signalling information being transferred. To measure this, we capture the validation BLEU score for Tr→En after 10k training steps.

In general, transferring from any monolingual or synthetic parent yields better BLEU (or faster convergence for Turkish) compared to training from scratch. Although, the improvement is suboptimal when compared with transfer learning from a proper parent. However, we can use these gains to measure the information transferred in transfer learning.

In general using monolingual English is better than using monolingual Chinese. In monolingual English, we can transfer the embedding information correctly with token matching. Therefore, consistent with our previous experiment, embedding information is transferred.

Using a Chinese parent is better than using random sequences. Our random sequence is uniformly sampled independently for each token. Therefore, unlike a real monolingual corpus, learning language modelling from this random sequence is impossible. Thus, we conclude that the model transfers some statistical properties of natural languages.

Transferring from a random sequence copy model yields better result compared to training the model from scratch. While the improvement is minimal, we can see that a naïve model that performs copying is better as a model initialisation. Moreover, substitution sequence parent models perform better than their copying counterparts. We suspect that copy models with tied embeddings converge to a local optimum that is a poorer initialisation for other translation models, compared to the substitution models.

Transfer learning with an actual NMT system as a parent still outperforms the monolingual and synthetic parents, albeit they are initially a copy model. We argue that the monolingual parents perform nearly perfectly at the copying task, and have perfect diagonal alignment, and therefore overfit to this artificial setting when used as a parent.

6 Transfer Learning for High-Resource Languages

Transfer learning can be used to initialise a model even if final quality does not change. Compared to random initialisation, we argue that a pre-trained model functions as better initialisation. Therefore, since we initialise the model better, it should converge faster. This behaviour was already presented in Table 7, where the transferred model converges more rapidly. However, we should explore this behaviour in a setting where faster training matters more: when training high-resource language pairs.

Parent	**Type**
Mono copy sequence (En→En)	**src:** Madam President , on a point of order . **tgt:** Madam President , on a point of order .
Mono substitution sequence (En$_S$ →En)	**src:** Click write , ideologies rotate sful ECHO recommended struggle **tgt:** Madam President , on a point of order .
Mono copy sequence (Zh→Zh)	**src:** 保持点神秘感。 **tgt:** 保持点神秘感。
Mono substitution sequence (Zh$_S$ →Zh)	**src:**比赛漂亮家宝1503 知识产权 **tgt:** 保持点神秘感。
Random copy sequence (Rand→Rand)	**src:** 1 3 2 1 1 **tgt:** 1 3 2 1 1
Random substitution sequence (Rand$_S$ →Rand)	**src:** 2 4 3 2 2 **tgt:** 1 3 2 1 1

Table 6: Monolingual and random parents with their sentence example.

	BLEU			
Parent	My→En	Id→En	Tr→En	Tr(10k)
-	4.0	20.6	19.0	14.3
De→En	17.8	27.4	20.3	20.2
En→En	10.4	23.3	18.5	16.0
En$_S$ →En	12.3	23.8	19.0	16.5
Zh→Zh	8.3	22.5	18.8	16.3
Zh$_S$ →Zh	11.2	23.5	19.0	16.3
Rnd→Rnd	6.2	21.9	19.0	15.2
Rnd$_S$ →Rnd	7.9	22.0	19.3	15.1

Table 7: Transfer learning performance on monolingual and synthetic parents. We also measure the validation BLEU of Tr→En after 10k updates.

Parent	BLEU	Num. Steps to 34 BLEU
Baseline	35.6	48k
+ no warm-up	0.0	-
En→En$_S$	35.4	60k (0.8x faster)
En→Ru	35.7	40k (1.2x faster)
+ token matching	35.7	34k (1.4x faster)
+ no warm-up	35.6	22k (2.1x faster)

Table 8: Transfer learning effect to the model's quality of high-resource language. We also measure the time to reach a near-convergence level of 34 BLEU.

For this experiment, we take an English-to-Russian model as a parent for an English-to-German model. We align the embedding with the same BPE tokens instead of using a joint vocabulary since this would require re-training the parent. We also attempt to exclude the embedding completely. These choices are practical in a real-world scenario, especially when we measure for efficiency.

In Table 8, we show that transfer learning does not improve the model's final quality. However, we can see both from the Table, and visually in Figure 2, that transfer learning speeds up the convergence by up to 1.4x, assuming the parent model has been prepared before.

In the early stage of training, the gradients produced are quite noisy, which is particularly harmful to the transformer model (Popel and Bojar, 2018). Therefore, training transformer models usually require a precise warm-up setup. However, transfer learning can be used as a better initialisation, thus skipping the noisy early training. To further confirm this, we remove the learning rate warm-up to observe the impact of a pre-trained model.

As shown in Figure 2, the pre-trained model remains capable of learning under more aggressive hyperparameters. On the other hand, the model without pre-training fails to learn. This result is congruent with the findings of Platanios et al. (2019), who found that warm-up in the Transformer can be removed with curriculum learning.

7 Conclusion

We demonstrate that the internal layers of the network are the most crucial for cross-lingual transfer learning. The embeddings contain transferable information, as long as the vectors are mapped correctly and the inner layers are also transferred. While not as optimal, we can still perform transfer learning by excluding the embedding. In transfer learning, we can also transfer the alignment. Transferred parents without fine-tuning will align

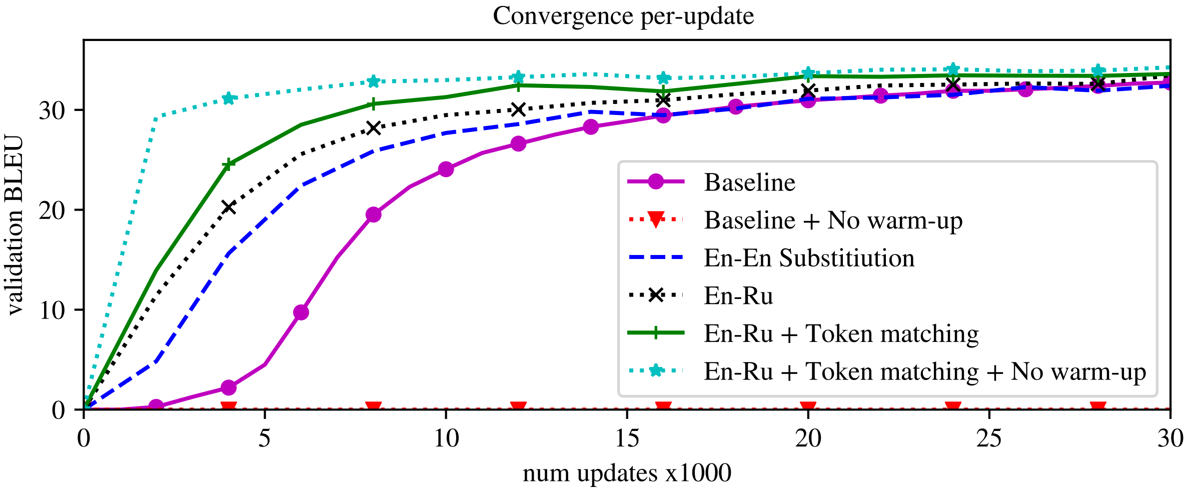


Figure 2: Transfer learning effect on the convergence of a high-resource system. Transfer learning removes the need for warm-up.

the input diagonally and copy most of the tokens. We further demonstrate that transfer learning still functions with a simple copy model, even with an artificial dataset—albeit with a reduced quality.

From a theoretical perspective, our results indicate that while transfer learning is effective in our scenario, it performed less "transfer" than previously thought. Therefore, a promising research direction to investigate would involve the development and assessment of improved initialisation methods that would more efficiently yield the benefits of the model transfer.

From a practical perspective, our results indicate that we can initialise models with a pre-trained model regardless of the parent language or vocabulary handling. With this perspective in mind, we can use transfer learning as a better initialisation, resulting in the child model having more stable gradients from the onset of training. Therefore, models can train and converge faster, which is useful in high-resource settings. With transfer learning, the model can be trained with more aggressive hyperparameters—such as removing the learning rate warm-up entirely—to further improve the convergence speed. This result further highlights the use of transfer learning as a better model initialisation.

Acknowledgments

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (http://www.csd3.cam.ac.uk/), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk). Alham Fikri Aji is funded by the Indonesia Endowment Fund for Education scholarship scheme. Rico Sennrich acknowledges support of the Swiss National Science Foundation (MUTAMUR; no. 176727).

References

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

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and

Wei Xu. 2016. Neural machine translation with pivot languages. *arXiv preprint arXiv:1611.04928*.

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.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. *on Language, Information and Computation*, page 282.

Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Towards Burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):5.

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2018. NOVA: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):17.

Mozhdeh Gheini and Jonathan May. 2019. A universal parent model for low-resource neural machine translation transfer. *arXiv preprint arXiv:1909.06516*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018a. 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)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018b. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. 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.

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, Brussels, Belgium. Association for Computational Linguistics.

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, Belgium, Brussels. Association for Computational Linguistics.

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.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

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)*, pages 296–301.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

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

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

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.

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*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.