A Study on Multilingual Transfer Learning in Neural Machine Translation: Finding the Balance Between Languages

Adrien Bardet, Fethi Bougares, and Loïc Barrault

LIUM, Le Mans University, France
firstname.lastname@univ-lemans.fr

Abstract. Transfer learning is an interesting approach to tackle the low resource languages machine translation problem. Transfer learning, as a machine learning algorithm, requires to make several choices such as selecting the training data and more particularly language pairs and their available quantity and quality. Other important choices must be made during the preprocessing step, like selecting data to learn subword units, the subsequent model's vocabulary. It is still unclear how to optimize this transfer. In this paper, we analyse the impact of such early choices on the performance of the systems. We show that systems performance are depending on quantity of available data and proximity of the involved languages as well as the protocol used to determined the subword units model and consequently the vocabulary. We also propose a multilingual approach to transfer learning involving a universal encoder. This multilingual approach is comparable to a multi-source transfer learning setup where the system learns from multiple languages before the transfer. We analyse subword units distribution across different languages and show that, once again, preprocessing choices impact systems overall performance.

Keywords: Transfer learning · Machine Translation · Languages proximity · Data quantity · Subwords distribution · Languages Balance · Data balance · Multilingual.

1 Introduction

Some major technical advances have allowed neural systems to become the most efficient approach to machine translation when a large amount of data is available [**?**,**?**]. However, when small amounts of training data are available, neural systems struggle to obtain good performance [**?**]. Transfer learning consists in training a first neural machine translation (NMT) system on another language pair where a larger quantity of data is available. This already trained system is then adapted to the new data from the low resource language pair with the aim of getting better performance than a system trained only on few data.

Transfer learning in machine translation results in learning a first system called "parent" system on abundant data. Then use this learned system as a basis for the "children" systems. This basis allows the system to learn the new data. It is comparable to domain adaptation where the domain is another language. The transfer generally improves the

results of the child system which benefit from knowledge learned during the training of the parent system [**?**]. In this paper we show that the data used to train the parent system significantly impact the resulting child system. Many factors must be taken into account like data quantity, languages proximity, data preprocessing, etc. Several studies deal with quantities of data used as well as proximity of involved languages, and conclusions diverge [**?**,**?**].

2 Related work

Currently, neural systems require a large body of training corpus to achieve good performance, which by definition is problematic for low resource language pairs. Phrase-based statistical machine translation systems appears then as a relevant alternative [**?**].

Several multilingual automatic translation approaches have been developed to translate texts from low resource language pairs [**?**,**?**,**?**]. The use of universal encoders and decoders allowed [**?**] to design a learning system that manages several language pairs in parallel and achieves better results, especially for less-endowed languages. Specific symbols (e.g. <2s>) are used to control the output language of the universal decoder. This kind of model even makes it possible to translate pairs of languages that are not seen during training (so-called *zero-shot learning*). However, performance in such cases remains relatively low. In the same line, [**?**] explore different parameter sharing schemes in a multilingual model.

The choice of level of representation of words is important. [**?**,**?**] have shown that the use of sub-lexical symbols shared between languages of the parent and child models results in an increase of transfer performance. In our work, we use this method by exploring different amounts of symbols (i.e. different vocabulary sizes).

Transfer learning tries to overcome the problem by relying on a parent system (trained on a large amount of data) that serves as the basis for learning a child system [**?**]. The transfer is more effective when languages are shared between the parent and the child. In this direction, [**?**] highlights the importance of proximity of languages in order to obtain a transfer of better quality. These observations are contradicted in [**?**] where better results are obtained with more distant but better endowed language pairs.

The work presented in this paper extends those of [**?**] and [**?**] on several points. Like [**?**], we try to evaluate the performance of the child system according to the data used in the parent system, still considering the criteria of proximity of language and quantity of data. In this study, we also consider a parent system consisting of a universal encoder (trained over several languages). We study different choices to be made for preprocessing data and the parameters of the translation model and we will try to determine the best configuration.

The objective is to better understand the correlation between the impact of the amount of data and the proximity of languages on the performance of the child system. We will see that our experiences contradict some of the conclusions of the articles mentioned above.

3 Data

In order to perform transfer learning, we need multiple language pairs. We selected the Estonian→English language pair as our low resource language pair. Our goal is to have the best possible results for this language pair, thus we need another language pair to train the parent system for the transfer.

3.1 Data selection

We use data provided in WMT2018 machine translation evaluation campaign [**?**]. 2.5 million (41M words source side and 52M target side) parallel sentences are available for the Estonian→English language pair. We can not consider this as a low resource language pair, but this quantity remains low for training an NMT system that achieves good results. To assess the impact of language proximity in the parent system, we use two pairs of different languages with different proximity to Estonian. The first pair is Finnish→English. Finnish is close to Estonian since both are Finno-Ugric languages. 5 million parallel sentences are available for this language pair which represent 78M words source side and 114M target side. As distant language pair we have chosen German→English. German is a Germanic language further away from Finnish and Estonian for which 40 million parallel sentences are available corresponding to 570M words source side and 607M target side. This will allow us to evaluate the impact of the quantity of data. Both pairs have English as target so the transfer will be from close (or distant) language on source side and target language is fixed. We want to exhibit whether this significant difference in terms of quantity of data will compensate for the language distance and result in a distant (German→English) parent system from which the transfer is as effective as a close (Finnish→English) parent system.

3.2 Data preprocessing

We use subword SPM units [**?**]. Systems using subword units are the current state of the art in neural machine translation. There is also a correlation in transfer quality depending on the number of subword units in common [**?**].

Two separate models of subword units are learned, one trained on source languages and another trained on the target language (English). Both are used in parent and child systems. The corresponding source and target vocabularies are created from tokenized data. Consequently, there is a direct correlation between data used to train the SPM models and resulting vocabularies in the NMT systems.

We take this into account by learning subword models for the source side with the data used to learn the parent **and** child systems. This subword model is then applied to all source side data, for both parent and child data. The goal is to not change the vocabulary during the parent/child transition since this would require to get representations for units that are not seen during training of the parent model.

Sentences with length less than 3 subword units and more than 100 subword units are filtered out.

Finally, subword units that occur at least 5 times in our training corpus are kept in the vocabularies while the others are matched to an unknown unit (<unk>). This

process is necessary in our case since SPM can not guarantee exhaustive coverage of the corpus.

3.3 SPM model study

In this section we will describe 3 different SPM models that were created for the multilingual parent system approach.

We tried to combine the proximity of Finnish and Estonian and to take advantage of large amount of data from the German→English pair. For that, we built a system with universal encoder and decoder [**?**] with both Finnish→English and German→English corpora used as training data. One advantage of the universal approach is the capability to add one or more languages to our system without having to change the architecture. We can thus always have really comparable Estonian→English children, whereas we now have a multilingual system as a parent. [**?**] showed that parallel learning of multiple language pairs with a universal architecture has a positive impact on translation results. We want to verify if this is also the case for transfer learning.

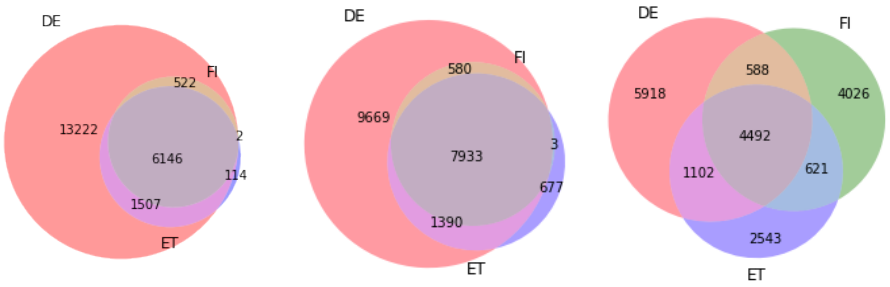

Fig. 1. Subword distribution for different SPM models built with different data distributions. Left is using 10M DE tokens only, center is using 5M DE tokens, 3M FI and 2.5M ET, and right is using 3M DE, 3M FI and 2.5M ET tokens.

We designed 3 different SPM data distributions to train separate SPM models for the source side of our universal systems.

For the first one, the model has been trained on German data only, resulting in subword units that are specific this language but are used for Finnish and Estonian. The result is that the vocabulary contains many short subword units covering the Finnish and the Estonian text. The Finnish and Estonian words end up being heavily split, which might complicate the subsequent modelling. This model is referred to as the 10-0-0 SPM. The second SPM model is using 5M German sentences, 3M in Finnish and 2.5M in Estonian. This is an intermediate model, with a more balanced data distribution across languages. It is referred to as the 5-3-2.5 SPM. The last one is made of 3M sentences in German, 3M in Finnish and 2.5M in Estonian. We force the data to be balanced for this SPM model despite the data quantity imbalance. It is referred to as the 3-3-2.5 SPM.

Figure 1 describes subword units distributions in the vocabulary of our systems obtained with the three different SPM models presented above. Figure 1 (left) shows the distribution for the 10-0-0 SPM model. This distribution is very unbalanced as expected. The Finnish specific units and Estonian specific units in this figure are unseen by the SPM model composed on German data. Figure 1 (center) shows that, even when greatly reducing the quantity of German data, the vocabulary remains mainly composed of German-specific units, however, we notice more common subword units than with the 10-0-0 SPM.

It is important to keep in mind that the distribution of subword units in the vocabulary does not reflect the actual coverage in the corpus (the number of occurrences is not taken into account).

The Figure 1 (right) shows a balanced distribution of subword units across the 3 languages.

In this case, every language has a set of specific units that will be learned during training of the parent and/or child NMT model. We want to verify whether this will lead to a better transfer for NMT.

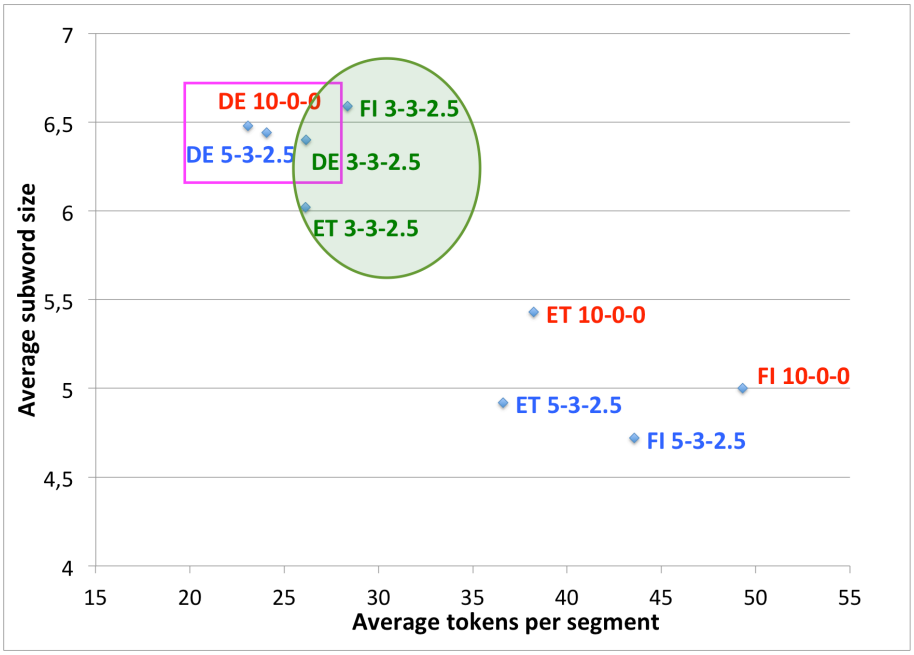

Fig. 2. Average subword units size (in characters) against average number of tokens per segment for the different SPM models.

Figure 2 shows the relationship between the average subword token size (in characters) and the number of tokens per segment. This graph highlights that all the ratios for German are close to each other regardless the data distribution used to train them (they are all in the pink rectangle). It also shows that the Finnish and Estonian ratios of each SPM model are close to each other in each distribution. There is a particularity observed only with the 3-3-2.5 SPM: all 3 languages ratio are close to each other (circled in green in Figure 2), emphasizing the balance between the 3 languages.

4 Architecture

To carry out our experiments we based our approach on the principle of trivial transfer learning of [**?**]. The principle is simple and consists in using an architecture that does not change between the learning of the parent system and the child system. Only training data are changed between parent and child system learning.

We use a standard end-to-end encoder/decoder architecture with an attention mechanism [**?**,**?**]. This architecture is composed of a bi-directional encoder and a decoder based on gated recurrent units, Bi-GRU [**?**] of size 800. The embeddings size is set to 400. We apply a 0.3 *dropout* [**?**] on the embeddings, on the context before being provided to the attention mechanism and on the output before softmax. Weights are initialized according to [**?**] and we use Adam [**?**] as optimizer. The initial learning rate is set to 1.10e-4 and size of a batch is 32. This architecture is the only configuration used for all systems presented in the next sections. They were implemented with the nmtpytorch[1] toolkit [**?**].

5 Experiments

All results of the systems presented here are calculated on the development dataset of the *News* translation task of WMT2018 evaluation campaign.

# subword units	ET-EN 2.5M	ET-EN 200k
8k	14.12	10.69
16k	14.17	10.70
32k	13.60	10.10

Table 1. Results in BLEU of the Estonian→English language pair without transfer learning with vocabularies containing only subword units coming from this language pair (source and target side separated).

The Estonian→English system presented in Table 1 is compared to our system using transfer learning. We can notice that the differences are small and negligible. Note that the number of source and target side subword units is the same.

[1] https://github.com/lium-lst/nmtpytorch

# subword units	DE+ET 2.5M	DE+ET 200k	FI+ET 2.5M	FI+ET 200k
8k	10.64	-	14.47	-
16k	11.55	9.27	15.08	10.66
32k	12.52	-	13.87	-

Table 2. Results in BLEU of Estonian→English systems without transfer learning using subword units from the different SPM models used in transfer learning afterwards. This emphasises the impact of the subword units and vocabulary used.

We can see in Table 2, that for learning a child system Estonian→English, the models based on the subword units including German get worse results than those including Finnish. Since Finnish and Estonian are close languages, it is likely that they share more subword units than with German, which explains the results. The hypothesis is that they coexist better in the vocabulary. This is confirmed by the results of the Estonian-German SPM model, which increases as number of subword units increases. While for the Estonian-Finnish SPM, results decrease when using 32k units compared to using 16k units. Therefore, it seems that a greater number of subword units is more favourable for German-Estonian system whereas 16k units are sufficient for the Finnish-Estonian system.

Language Pair	40M	20M	10M	5M	2.5M
FI-EN	-	-	-	18.03	16.16
DE-EN	20.22	10.46	11.01	11.11	10.95

Table 3. Results in BLEU of different parent models in their respective languages pairs.

The results of the standard systems in Table 3 give us an idea of the performance obtained by parent systems in their respective language pairs. We selected 5M sentences from the German→English corpus to have a similar size to the Finnish→English corpus. We can then effectively compare the 2 different source languages of the parent systems with the same amount of data to train on. This allows us to evaluate the impact of language proximity. We observed that the performance of the German→English parent system using only 5M of randomly selected data is much lower than that of the system using all available data. One can thus expect a loss of performance when the parent system is trained with a smaller amount of data.

We had to make a compromise when defining architecture size. We want the biggest possible architecture to effectively learn the parent system, but we also want a reasonable size to avoid overfitting the child's system afterwards. For the upcoming experiments, we chose to use 16k subword units because this quantity led to the best performance for the Estonian→English system.

5.1 Results

In Table 4, we expose the results of the Estonian→English child systems that were learned from the different parent systems presented in Table 3.

Using all data for our 3 systems we get an improvement compared to our Estonian→English baseline which is 14.17 BLEU (see Table 1). Results are close but we can see that the results with the DE+ET SPM are worse, which corresponds to results in Table 2. The best result is obtained with the Finnish→English parent.

We performed several experiments with different quantities of German→English data (from 40M to 1.25M) and Finnish→English data (from 5M to 1.25M). Results show that with the same amount of data, results differ greatly among architectures. This difference is explained by the proximity of languages used to train the parent system. Finnish, which is closer to Estonian, offers a better transfer than the more distant German, confirming results in [**?**]. [**?**] shows that the quality of the parent system is important to ensure a good transfer to a child. The low performance of the German→English parent using 5M of data explains the poor results of the later learned child system.

We also tried our multilingual parent approach with universal encoder as described in Section 3.3.

We use a different SPM model from previous ones because this time it contains German and Finnish from the parent system, in addition to Estonian from the child system for source side.

The assumption is that by combining these two factors we should get a parent who will provide a better transfer to our child systems. The results show that this is not so obvious (see Table 4); performance is worse than German→English or Finnish→English as the only parent. One hypothesis is that the imbalance of amounts of data between the two source languages of the parent is an obstacle to learning a good quality parent.

Parent Language Pair	45M (40M DE + 5M FI)	40M	20M	10M	5M	2.5M
FI-EN	-	-	-	-	**16.55**	**16.55**
DE-EN	-	**16.10**	10.46	11.28	10.92	11.18
FI+DE-EN 10-0-0 SPM	**15.71**	14.06	14.44	14.37	14.53	14.47
FI+DE-EN 5-3-2.5 SPM	13.20	13.45	13.86	14.05	14.01	13.77
FI+DE-EN 3-3-2.5 SPM	14.09	14.51	14.22	14.64	14.52	14.71

Table 4. Results in BLEU of Estonian→English child models with different parent models used to transfer.

We tried different data quantity distribution between the two languages.

In the 40M column there is 35M from German→English with 5M from Finnish→English, in the 20M column it is 15M German→English with 5M Finnish→English, in the 10M column it is 5M German→English with 5M Finnish→English, in the 5M column it is 2.5M form both and in the last column it is

1.25M from both. This way we have a better vision of the impact of data quantity from the German→English language pair.

With our 3 different SPM models applied to our multilingual parent, we observe some interesting results.

The first of them is the 15.71 BLEU obtained when using the full data available on both parent pairs.

This result is surprising considering the others scores. This SPM model learned only on German→English data reveals an interesting behaviour of non DE source side data. Indeed, words are "overly" split into subword units. Surprisingly, this particularity seems to provide good results in Estonian→English. Our hypothesis is that, thanks to the quite large architecture to train the Estonian→English system, the system overfits on the small subword units. The average number of subword units per line and average subword unit size of this particular SPM model applied to our data can confirm it. Thus, for the training of this child model, the system keeps improving and stopped during the 10th epoch, while most of the other child systems presented barely passed the 5th. These observations are in line with our hypothesis of overfiting but we keep investigating about this result.

Overall the results from the 5-3-2.5 SPM are the least interesting. Results seem to increase slightly as we reduced data quantity used by a small margin. This might be related to the quantity of German→English data used for training the model.

Finally, with the more balanced SPM model (3-3-2.5), the results are quite stable with only slight changes. The results are also better at each data quantity than the 5-3-2.5 SPM.

We see an improvement thanks to the transfer for the Estonian→English child systems. However, we also want to apply this transfer in the case where few resources are available. To simulate this lack of data, we kept only 200k sentences from the Estonian→English corpus to learn new children with the same parent systems.

Parent Language Pair	45M (40M DE + 5M FI)	40M	20M	10M	5M	2.5M
FI-EN	-	-	-	-	**13.03**	**12.24**
DE-EN	-	**11.12**	6.87	6.99	7.10	6.96
FI+DE-EN 10-0-0 SPM	**11.05**	10.41	11.29	11.68	11.54	11.72
FI+DE-EN 5-3-2.5 SPM	10.26	9.79	10.52	11.00	10.85	10.65
FI+DE-EN 3-3-2.5 SPM	**12.19**	12.05	11.89	**12.56**	11.93	**12.45**

Table 5. Results in BLEU of the Estonian→English language pair using only 200k sentence pairs to train the child model (artificially simulated low ressourced)

Results of Table 5 show us that when we have few training data for the child system, the proximity of languages is the most important feature.

Finnish→English parent system outperforms the others in this configuration.

This time our multilingual approach results are as good as with the German→English pair. Results coming from the balanced SPM outperforms it on all data quantities.

Compared to previous results with all the data on the 10-0-0 SPM system, we observe that the results are not anymore outperforming the others. The system obtains 11.05 BLEU which is not better than the previous results. This can confirm that the overfiting of the subword units works less well when less data is available like in this setup. In general the results are consistent without dependency to data quantity.

With the 5-3-2.5 SPM model, the results are consistent as before: they are still worse than the two others SPM models.

Our 3-3-2.5 SPM models now outperforms the others as well as the German→English parent systems. Their results even get close to the Finnish→English parent. We believe that the balance of the subword units across the 3 source languages involved is particularly effective in this case where few data are available for the child system.

6 Conclusion

In this paper, we showed that transfer learning for NMT depends on the quantity of available data and the proximity of involved languages. Also, carefully training the subword models can lead to a better language equilibrium in the vocabulary leading to better translation results.

These parameters are therefore to be taken into account for the choice of parent systems.

Our results are in line with those obtained by [**?**] and [**?**]; proximity of languages used for transfer learning is more important than data quantities. With equivalent amounts of data, parent systems using pairs of closer languages perform better, but the quality of the parent systems in question should not be neglected and should be taken into account in the results of the child systems. The token distribution in the vocabulary is also of greater importance and have an impact on system performance.

Our universal multilingual approach end up showing some interesting results, especially in low resource context. We presented an analysis of the subword units distribution and the importance of the preprocessing steps ahead of the training process. We showed that the balance between the different languages involved in the system is extremely relevant for the performance of the child systems afterwards.

In the future we want to keep investigating subword units distribution with different examples to better explain the relation between those factors and the systems performance results.

Acknowledgments

This work was supported by the French National Research Agency (ANR) through the CHIST-ERA M2CR project, under the contract number ANR-15-CHR2-0006-017.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR **abs/1409.0473** (2014), `http://arxiv.org/abs/1409.0473`

2. Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Monz, C.: Findings of the 2018 conference on machine translation (wmt18). In: Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers. pp. 272–307. Association for Computational Linguistics, Belgium, Brussels (October 2018), http://www.aclweb.org/anthology/W18-6401
3. Caglayan, O., García-Martínez, M., Bardet, A., Aransa, W., Bougares, F., Barrault, L.: Nmtpy: A flexible toolkit for advanced neural machine translation systems. Prague Bull. Math. Linguistics **109**, 15–28 (2017). https://doi.org/10.1515/pralin-2017-0035, https://ufal.mff.cuni.cz/pbml/109/art-caglayan-et-al.pdf
4. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR **abs/1406.1078** (2014), http://arxiv.org/abs/1406.1078
5. Dabre, R., Nakagawa, T., Kazawa, H.: An empirical study of language relatedness for transfer learning in neural machine translation. In: Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation. pp. 282–286. The National University (Phillippines) (2017), http://aclweb.org/anthology/Y17-1038
6. Durrani, N., Dalvi, F., Sajjad, H., Belinkov, Y., Nakov, P.: One size does not fit all: Comparing NMT representations of different granularities. 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). pp. 1504–1516. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://www.aclweb.org/anthology/N19-1154
7. Gu, J., Wang, Y., Chen, Y., Li, V.O.K., Cho, K.: Meta-learning for low-resource neural machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3622–3631. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018), https://www.aclweb.org/anthology/D18-1398
8. Ha, T., Niehues, J., Waibel, A.H.: Toward multilingual neural machine translation with universal encoder and decoder. CoRR **abs/1611.04798** (2016), http://arxiv.org/abs/1611.04798
9. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. CoRR **abs/1502.01852** (2015), http://arxiv.org/abs/1502.01852
10. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F.B., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's multilingual neural machine translation system: Enabling zero-shot translation. CoRR **abs/1611.04558** (2016), http://arxiv.org/abs/1611.04558
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://arxiv.org/abs/1412.6980
12. Kocmi, T., Bojar, O.: Trivial transfer learning for low-resource neural machine translation. In: Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018. pp. 244–252 (2018), https://aclanthology.info/papers/W18-6325/w18-6325
13. Koehn, P., Knowles, R.: Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation. pp. 28–39. Association for Computational Linguistics, Vancouver (Aug 2017). https://doi.org/10.18653/v1/W17-3204, https://www.aclweb.org/anthology/W17-3204
14. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. CoRR **abs/1808.06226** (2018), http://arxiv.org/abs/1808.06226
15. Lakew, S.M., Erofeeva, A., Negri, M., Federico, M., Turchi, M.: Transfer learning in multilingual neural machine translation with dynamic vocabulary. In: IWSLT'18 (10 2018)

16. Nguyen, T.Q., Chiang, D.: Transfer learning across low-resource, related languages for neural machine translation. CoRR **abs/1708.09803** (2017), http://arxiv.org/abs/1708.09803
17. Sachan, D., Neubig, G.: Parameter sharing methods for multilingual self-attentional translation models. In: 3nd Conference on Machine Translation (WMT). Brussels, Belgium (October 2018), https://arxiv.org/abs/1809.00252
18. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014), http://jmlr.org/papers/v15/srivastava14a.html
19. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. CoRR **abs/1409.3215** (2014), http://arxiv.org/abs/1409.3215
20. Zoph, B., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. CoRR **abs/1604.02201** (2016), http://arxiv.org/abs/1604.02201