Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 435–446
https://doi.org/10.22364/bjmc.2022.10.3.16

Cross-lingual Transfer from Large Multilingual Translation Models to Unseen Under-resourced Languages

Maali TARS, Andre TÄTTAR, Mark FISHEL

University of Tartu, Ülikooli 18, 50090 Tartu, Estonia

maali.tars@ut.ee, andre.tattar@ut.ee, mark.fisel@ut.ee

Abstract. Low-resource machine translation has been a challenging problem to solve, with the lack of data being a big obstacle in producing good quality neural machine translation (NMT) systems. However, recent work on developing large multilingual translation models gives a platform for attempts to create NMT systems for extremely low-resource languages that can achieve reasonable and usable quality. We leverage the information in large multilingual translation models by performing cross-lingual transfer learning to extremely low-resource Finno-Ugric languages. Our experiments include seven languages with limited resources that are unseen by the original pre-trained translation model and five high-resource languages that have the potential to help during training, previously seen by the model during training. We report state-of-the-art results on multiple test sets and translation directions as well as analyze the low-resource languages in smaller language groups in order to track the source of our higher translation quality.

Keywords: multilingual, cross-lingual transfer learning, low-resource, Finno-Ugric languages

1 Introduction

In previous years, there has been a gap between bilingual and multilingual neural machine translation (NMT) models in terms of translation quality (Johnson et al., 2017). Lately, however, the multilingual translation model quality has improved a lot (Aharoni et al., 2019; Zhang et al., 2020), resulting in multilingual translation models being a preferred choice over bilingual models, as it enables translations in multiple directions, while only training and deploying one model.

For low-resource languages, it has been shown that the experiments achieve much better results in the multilingual setting than in the bilingual setting (Johnson et al., 2017; Gu et al., 2018; Rikters et al., 2022). The reasoning behind it is that the low-resource language pairs often have very little data available to train an NMT model,

but the multilingual setting enables them to leverage the information provided by the high- and medium-resource language pairs that are included in the same multilingual translation model.

In our task, we apply the technique of doing cross-lingual transfer learning to the languages that were previously unseen by the pre-trained model. Our focus in this paper is on the Finno-Ugric language family and we work with languages that are extremely low-resource: Livonian (liv), Võro (vro), North Sami (sme), South Sami (sma), Inari Sami (smn), Lule Sami (smj), and Skolt Sami (sms).

Recently there has been valuable research into creating large-scale multilingual machine translation models trained on data from one hundred or more languages (Fan et al., 2021). These models have achieved state-of-the-art translation results on multiple benchmarks for various translation directions, mainly for non-English language pairs on various resource levels. This is promising for low-resource languages because of how much information the large multilingual models contain, making them a reasonable starting point when trying to improve translation quality for extremely low-resource languages. Our main contributions are the following:

- Training the first NMT models for the low-resource Finno-Ugric languages of Inari Sami, Lule Sami, and Skolt Sami.
- Identifying problems in existing test data and creating a new benchmark dataset for Finno-Ugric language pairs, which are publicly shared.
- Providing an example of performing cross-lingual transfer learning on large pre-trained NMT models to languages not included in the initial training of the model, using the example of low-resource Finno-Ugric languages.
- Reporting state-of-the-art results for low-resource Finno-Ugric language pairs as well as analysis of what were the factors that helped enhance the translation quality.

Firstly, we are going to describe the technical steps needed to start executing transfer learning on large pre-trained multilingual translation models to unseen languages. In the next section, we present the data we used to train the models and give details on the new test dataset, including the description of the pre-processing and filtering process. This is followed by setting up all of the experiments and subsequently analyzing the results. We conclude by outlining the plan for future efforts in the field of NMT for low-resource languages.

2 Related Work

2.1 Low-resource NMT

The topic of transfer learning for low-resource languages has been previously studied by Gu et al. (2018), who introduce a universal NMT setting, where languages in the encoder share lexical and sentence level information with each other. In this setting, the languages exploit the word- and sentence-level similarities between the languages with a model of language experts.

Another work that focuses on finding solutions to the low-resource setting problem is by Sennrich and Zhang (2019), where they analyze different practices and offer insight into what might be the best technique to use in these circumstances. They note

that a lot of low-resource task solutions have been through data augmentation and multilinguality. In their experiments, however, they also mark that models trained on less data are very sensitive to hyperparameter and vocabulary size changes, which indicates that tuning these aspects could be beneficial.

Moving into the family of Finno-Ugric languages, Rikters et al. (2018) show that for Estonian, the multilingual setting is preferred, but with the cost of lowering translation quality between high-resource languages involved in the training of the multilingual model. Kocmi and Bojar (2018) dealt with bilingual models but emphasized how useful transfer learning can be for low-resource languages. They train a "parent" model on a high-resource language pair and then further train it on the desired low-resource language pair data. In their work they had two settings for the parent model: 1) includes one related language for the low-resource language, 2) does not include any related high-resource languages. In some cases, they report achieving higher quality with the second setting, which could be an indication that very different high-resource languages might even provide more useful information during training.

In the field of low-resource Finno-Ugric languages, it has previously been shown in Tars et al. (2021) that multilingual training with high- or medium-resource languages is very beneficial to languages with almost no parallel data available, like Võro, North Sami, and South Sami. The article also notes that the intuition to train similar languages together, as they might help each other, yields good results, e.g., using Finnish to improve Estonian-Võro translation. In this current work, we widen the language selection by including Livonian and three other Sami languages: Inari Sami, Lule Sami, and Skolt Sami.

For Livonian, we base our work on the results achieved by Rikters et al. (2022). Their experiments included an analysis of which base language is suitable for cross-lingual transfer for Livonian by pre-training a multilingual translation model on related medium to high-resource languages, Estonian, Latvian, and English. They conclude that using a multilingual model gets the best results, but Estonian helps Livonian the most.

With the emergence of various large pre-trained multilingual translation models, we can perform experiments which leverage more information than previous works.

2.2 M2M-100

In our work, we use Facebook's M2M-100 model (Fan et al., 2021). Previous approaches have tried to increase the quality of multilingual models by increasing the model's capacity and leveraging only English-centric data, leaving non-English-centric directions behind (Aharoni et al., 2019; Zhang et al., 2020). Fan et al. (2021) show that with their approach, training non-English-centric directions achieves better results than bilingual models when trained as part of the same model with English-centric language directions.

Using a large dataset for 100 languages means that the model's capacity to learn should also grow from the standard sizes, in order to not underfit the model. They solve this by adding a layer of parallel language-specific layers for language-specific parameters and also implementing a re-routing scheme between the said language-specific layers into their architecture to increase low-resource and high-resource languages sharing

information. As a result, their models, which are based on the Transformer architecture (Vaswani et al., 2017), are scalable and achieve state-of-the-art results on different benchmarks for non-English-centric language pairs, with an average increase of 10 BLEU points in translation quality.

With regards to the low-resource setting, they make efforts to create a balanced dictionary between languages of all resource levels and also upsampling low-resource data shards to have an equal capacity to high-resource language pairs. Among their analysis, they find that having language-specific parallel layers and a re-routing scheme is very beneficial, especially for low-resource languages. They conclude with a thought, which creates a premise for our work, that very low-resource language translation is still a problem due to the lack of data, and suggest ways to obtain more data.

We try to improve on the previous efforts made for extremely low-resource Finno-Ugric languages by leveraging this new large multilingual translation model by doing cross-lingual transfer learning.

3 M2M-100 enhancement

For training the M2M-100 model, the authors chose languages that had an existing evaluation benchmark and a significant amount of monolingual data available. Additionally, their objective was to cover different language families and languages of various resource levels. However, extremely low-resource languages, like the small Finno-Ugric languages used in our work, often do not fill those criteria, which means they were not included in the training process.

One possible approach to achieve better translation quality for smaller language pairs included in the M2M-100 model would be to fine-tune it on a specific language pair. We are, however, attempting to do cross-lingual transfer to new unseen languages. Adding new languages requires changing the embedding matrix and possibly having to introduce new tokens to the vocabulary (additional to the new language ID tokens), in order to avoid producing texts with a high percentage of unknown symbols. Our data includes multiple languages that have tokens unique only to these languages and are thus not known to the M2M-100 tokenizer.

3.1 Adding new tokens

There are two types of token changes we have to implement: 1) adding new language ID tokens, 2) adding new symbols that would otherwise cause `UNK` tokens in the translation. We make use of the HuggingFace implementation of M2M-100 tokenizer and add new language tokens as special tokens. The new language ID tokens are then associated with their ID-s in the vocabulary, and each language code is mapped to its token form.

For the additional language-specific symbols, we have to increase the embedding matrix size of both the encoder and decoder as well as give them indexes in the vocabulary by increasing the largest index previously present in the vocabulary and initializing the token vectors randomly. The code is available online.[1]

[1] https://github.com/TartuNLP/m2m-100-finetune

4 Data

4.1 Low-resource Finno-Ugric language data

We gathered data for various language pairs that are in the Finno-Ugric family or are connected to the smaller Finno-Ugric languages that might help the translation process. Building on the work done by Rikters et al. (2022) and Tars et al. (2021), we expand the language selection by three Sami languages: Inari Sami, Lule Sami, and Skolt Sami. We chose these languages because they had publicly available parallel data, more specifically pairings with North Sami, South Sami and two high-resource languages, Finnish and Norwegian (Bokmål).

Overall, our experiments include seven extremely low-resource Finno-Ugric languages (Livonian (`liv`), Võro (`vro`), North Sami (`sme`), South Sami (`sma`), Inari Sami (`smn`), Lule Sami (`smj`), Skolt Sami (`sms`)) and five high- or medium-resource languages connected to the smaller ones: Finnish (`fi`), Estonian (`et`), Latvian (`lv`), Norwegian (`no`), and English (`en`). Finnish and Estonian belong to the Finno-Ugric family, but other languages were included because they have existing parallel data with the smaller Finno-Ugric languages or are geographically and orthographically close. In Table 1, we can see the amounts of data for each language pair between the mentioned languages that we managed to gather. For the language pairs between the high- and medium-resource languages, we sample data from corpora obtained from OPUS (Tiedemann, 2012).

The Estonian-Võro (`et-vro`) data is mainly from parallel sentences of various domains acquired from META-SHARE[2]. For Livonian, the data is from a publicly available corpus named liv4ever in OPUS, curated by Rikters et al. (2022). The data is also diverse, ranging from excerpts from Facebook posts to the Latvian constitution document. Parallel data which includes any of the Sami languages was collected from publicly available translation memory files[3] compiled by The Arctic University of Norway.

4.2 Pre-processing and filtering

Detokenization and normalization of the data were done with Moses scripts. The normalization script was slightly altered by leaving out the language-specific conditions. Filtering was done with the OpusFilter tool (Aulamo et al., 2020). The basic filters are a slight modification of OpusFilter's default settings, because some thresholds had to be adjusted in order to remove more noise. The filters included (i) maximum word length (50), (ii) maximum length of segment (1000 chars, 400 words), (iii) difference in source and target segment length (3 times), (iv) ratio of alphabetic characters (0.75 or more), (v) ratio of characters in the correct alphabet, (vi) ratio of numerals in the sentence (0.5 or less).

Before training, the data was tokenized by the enhanced tokenizer with added symbols for the new Finno-Ugric languages. The tokenizer in the HuggingFace implementation uses SentencePiece (Kudo and Richardson, 2018). To signal to the model in which

[2] https://doi.org/10.15155/1-00-0000-0000-0000-001A0L
[3] https://giellalt.uit.no/tm/TranslationMemory.html

lang-pair	raw	filtered
et-vro	31 551	29 775
fi-sme	77 710	62 837
fi-sma	2913	2766
fi-smn	10 639	9459
fi-sms	5769	2708
no-sma	17 388	15 702
no-sme	241 598	195 970
no-smj	12 400	11 627
sme-sma	21 993	19 963
sme-smj	16 440	14 985
sme-smn	934	894
en-liv	617	280
et-liv	14 261	12 887
lv-liv	11 732	10 763

Table 1: Parallel data numbers before and after filtering (in sentence pairs).

lang-pair	test	valid
et-vro	500	200
fi-sme	500	200
fi-sma	500	200
fi-smn	500	200
fi-sms	500	200
no-sma	500	200
no-sme	500	200
no-smj	500	200
sme-sma	500	200
sme-smj	500	200
sme-smn	500	200
en-liv	856	586
et-liv	856	586
lv-liv	856	586

Table 2: Evaluation and validation datasets (in sentence pairs).

direction we are currently training the model, we add a language token in front of each sentence in the source as well as the target side; in inference mode the target-side token is forced instead of being predicted (this follows the original M2M-100 approach).

4.3 Evaluation and validation data

For all of the translation directions we train for, we extract new held-out validation and evaluation data[4] from the filtered training dataset, except for Livonian, for which there exists a ready-made benchmark (Rikters et al., 2022). In addition, we evaluated language pairs involved in Tars et al. (2021) on the same test data that was used in that article. An overview of the evaluation and validation data quantities is shown in Table 2.

However, our analysis found several problems with both of the cited datasets. Firstly, the Livonian dataset suffers from leakage, with 97 parallel training sentences appearing in the test set. To combat leakage, we removed these sentences from the training data.

Secondly, the Finno-Ugric dataset by Tars et al. (2021) used a held-out dataset without filtering and suffers from low-quality sentences appearing in the test set. After filtering and careful overlap checks with whitespace and punctuation removed, we created a new held-out test set. We compare our models on both the new and the cited test data.

5 Experiments

We perform multiple experiments with different datasets and on two different M2M-100 model sizes: 418 million parameters (418M) and 1.2 billion parameters (1.2B).

[4] https://huggingface.co/datasets/tartuNLP/finno-ugric-benchmark

One of the main comparisons is between the learning curves of the smaller and the larger M2M-100 models in cross-lingual transfer setting to our selected Finno-Ugric languages. Additionally, we train a number of other models on the smaller, 418 million parameter M2M-100 model. In this case, we leave some of the parallel data out in each of the experiments, according to the language group that we want to inspect in more detail.

5.1 M2M-100 418M vs 1.2B

The aim of comparing two different-sized M2M-100 models, was to see the learning curve of both of the models in the low-resource fine-tuning setting. Specifically, whether at some point in the training, the smaller model would reach the same level as the larger model or not. In this experiment, we added sampled data for the language pair directions we had between high- and medium-resource languages (`fi, et, no, lv, en`) in order to help the transfer learning and avoid forgetting the high-resource language pairs. For each of those translation directions, we sampled 20 000 sentences to balance them with the amount of low-resource language data.

5.2 Removing English

English is not a close language to any of the low-resource languages that we are including in our experiments. For that reason, we investigate whether it helps or rather takes up too much room in the training space. In order to do that, we remove all parallel data directions from the entire dataset that have English as one language of the pair.

5.3 Language groups inside Finno-Ugric language family

For finding out how languages affect each other during the training, we conduct multiple smaller experiments. We can group the Finno-Ugric languages that we use into smaller groups, based on geographical location and other Finno-Ugric languages that they are more similar to. In our experiments, we can separate Livonian, Võro and Sami languages into different models. For Võro, we only had Võro-Estonian data, so for the separate experiment the transfer learning takes place only on that language pair. For Livonian, we have parallel data directions `liv-et`, `liv-en`, `liv-lv` and all the data between `et-en-lv`. For Sami languages, the high-resource languages that they have parallel data with are Finnish and Norwegian. So the language pairs that we have are between `fi-no-sme-sma-smj-sms-smn`.

It is worth noting that in the case of Livonian, we include also Latvian data. Although Latvian is not a Finno-Ugric language, Livonian and Latvian have co-existed for centuries and have thus influenced each other's development. In addition, since it is spoken in Latvia, there is more public parallel data available for the `liv-lv` pair than with any other language. A similar case goes also for the Sami languages. We include Norwegian, because a lot of the Sami language speaking communities live in Norway and are thus influenced by it, primarily orthographically. Some of the special symbols are similar and a large portion of the parallel data that is available is paired with Norwegian.

5.4 Technical setup

We trained all our models on one Tesla a100 GPU with 40GB vRAM. For training we used the HuggingFace implementation of Transformers and the M2M-100 models that are available there[5]. The smaller model (418 million parameters) has 12 encoder and decoder layers, a feed-forward network of size 4096, 16 attention heads in each layer and an embedding size of 1024. The larger model (1.2 billion parameters) has 24 decoder and encoder layers, feed-forward network of size 8192. We initialize all our models with the default learning rate of HuggingFace code, which is 5-e05. Batch size was set to 12 with gradient accumulation steps set to 8. We trained all our models for 25 epochs.

6 Results

6.1 M2M-100 418M vs 1.2B

We did transfer learning on two sizes of the M2M-100 model with all of the data. We trained both model sizes for 25 epochs. In Figure 1, we can see the comparison of the learning progress on validation data averaged over all the language pairs mentioned in the data table. We can note that the two models seem to on average increase in quality at the same rate, but in the first 25 epochs the 418 million parameter model does not catch up to the 1.2 billion parameter model.

However, when inspecting the translation directions one-by-one, we see that for some of them the smaller model does catch up during the 25 epochs of transfer learning and for some the performance of 1.2 billion parameter model actually gets worse. This can be seen in Figure 2.

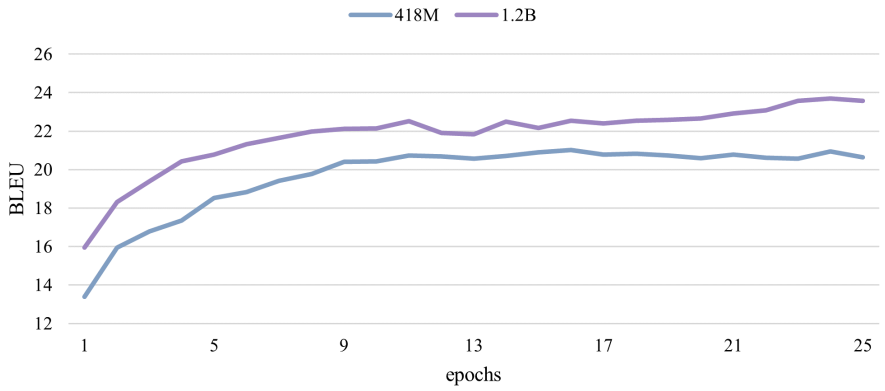
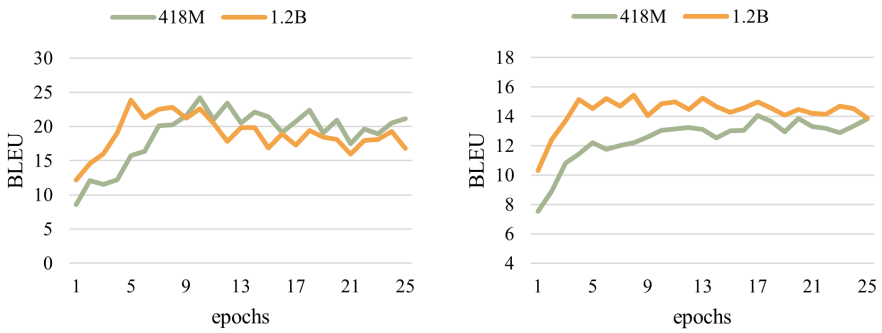


Fig. 1: 418M vs 1.2B model on validation data for 25 epochs (averaged over all low-resource language pairs).

[5] https://huggingface.co/docs/transformers/model_doc/m2m_100



(a) 418M vs 1.2B model for `no-sma`.

(b) 418M vs 1.2B model for `et-liv`.

Fig. 2: Comparison of 418M vs 1.2B models on specific language pairs.

For any conclusive results, the models should be trained longer because we cannot say if overall the learning progress has converged or not. This would be a part of future research.

6.2 Results on our new test data

The first thing we notice when looking at the results in Table 3 and 4 is that the 1.2 billion parameter model generally outperforms any other model, except for the Võro and Skolt Sami translation directions. Keeping in mind that we trained all the models for 25 epochs, it is an impressive difference of potential between the different-sized models for doing transfer learning.

The larger model was trained with low-resource data and sampled high-resource data. Comparing the 418 million parameter models (columns "418M" and "only-*" in Table 3) we note that putting all of the data together does not work as well as dividing them into their smaller language groups. The biggest gain from training in smaller specific groups is achieved for the Võro-Estonian language pair with an average of 4 BLEU points difference and for a couple of Sami translation directions `sma-fi`, `sms-fi`.

Another interesting analysis point is that removing English from data did not seem to have any real effect on the results, except of course for the `en-liv` pairing. Comparing the BLEU scores between the 418M model trained on all data and the model with no English shows less than 1 BLEU point difference on average.

6.3 Comparison to previous research

We had a chance to compare to some of the earlier neural machine translation quality scores for languages like Livonian, Võro, North and South Sami. For Livonian we evaluated the models on test data from Rikters et al. (2022) and compared to the results they report after tuning the baseline model to a specific language pair. The results can be seen in Table 3. While the smaller sized models achieved somewhat similar results

Model	418M	not-en	only-(vro/liv)	1.2B	*prev_best*
et-vro	29.28	30.38	**34.11**	30.06	-
vro-et	36.29	35.39	**40.04**	37.16	-
en-liv	10.64	2.51	8.97	**11.51**	*8.59*
liv-en	14.08	4.34	14.32	**15.85**	*14.69*
et-liv	13.97	13.7	13.87	**14.51**	*13.00*
liv-et	19.03	18.3	18.69	**19.62**	*17.76*
lv-liv	13.64	13.43	13.89	**15.13**	*13.67*
liv-lv	18.39	18.34	20.34	**20.47**	*17.55*

Model	418M	not-en	only-sm	1.2B
fi-sme	40.19	40.18	41.28	**42.89**
sme-fi	48.51	47.87	47.72	**50.14**
fi-sma	17.43	18.15	21.72	**26.63**
sma-fi	20.84	22.22	27.91	**31.45**
fi-smn	52.11	52.14	53.15	**53.3**
smn-fi	70.57	70.99	74.27	**75.43**
fi-sms	32.88	32.63	**33.72**	33.13
sms-fi	56.97	58.9	**61.49**	61.39

Model	418M	not-en	only-sm	1.2B
no-sma	43.94	43.25	45.9	**46.79**
sma-no	50.42	49.92	51.34	**53.52**
no-sme	34.56	**35.38**	35.19	34.93
sme-no	44.75	44.8	44.95	**45.84**
no-smj	33.42	34.39	36.98	**40.01**
smj-no	48.16	48.17	49.95	**52.42**

Model	418M	not-en	only-sm	1.2B
sme-sma	33.12	33.7	35.64	**36.32**
sma-sme	37.75	37.53	40.27	**44.41**
sme-smj	28.71	31.06	31.14	**34.06**
smj-sme	38.91	39.88	42.48	**46.41**
sme-smn	27.38	28.52	29.95	**33.54**
smn-sme	32.63	30.54	31.73	**34.21**

Table 3: BLEU scores on our new test set and liv4ever test set. "418M" and "1.2B" refer to models trained with all data. "not-en" refers to model trained without any English data. "only-(vro/liv/sm)" refers to a models trained only on that specific language group data. *prev_best* refers to results by Rikters et al. (2022) on their fine-tuned models (without back-translation). **bold** - best BLEU score for a language pair.

compared to the previous tuned results by Rikters et al. (2022), the 1.2 billion parameter model clearly achieves better translation quality for all of the Livonian translation directions.

As for Võro, North and South Sami, all of our models also improve in quality over the previous results reported in Tars et al. (2021), which can be seen in Table 4. For Võro, the best model is again trained on only Estonian-Võro data. This indicates that our newly created test set and the test set used in Tars et al. (2021) agree on which is the best model. For the Sami languages, we can see that the gain in BLEU is very significant, with sma-sme jumping about 22 BLEU points and the overall average gain being at 14 BLEU points. This is a massive leap forward from previous best results, which were reported after two iterations of back-translation. In this work, however, we did not utilize any monolingual data, which indicates how powerful the large multilingual translation systems are.

7 Future Work

As mentioned, in our current work, we did not gather any monolingual data, but creating synthetic data and enhancing the models with back-translation is a direction worth exploring. Additionally, the model size of M2M-100 could make the models inconvenient to train and subsequently to deploy, which could be eased by reducing the size of the embeddings, removing unnecessary alphabets in M2M-100 that do not overlap with the transfer-learned languages. Including other Finno-Ugric languages in the fine-tuning

Model	418M	not-en	only-(vro/sm)	1.2B	*prev_best*
et-vro	25.72	25.73	**30.32**	26.03	*26.2*
vro-et	30.27	29.66	**33.95**	31.65	*31.7*
fi-sme	**38.45**	38.3	38.02	37.83	*32.3*
sme-fi	42.84	45.09	45.21	**45.83**	*37.5*
fi-sma	17.82	20	22.56	**25.29**	*12.4*
sma-fi	21.73	21.89	27.85	**29.44**	*10.9*
sme-sma	33.54	33.74	34.93	**38.1**	*21.6*
sma-sme	35.44	36.54	38.37	**42.97**	*21.0*

Table 4: BLEU scores with test data from Tars et al. (2021). "418M" and "1.2B" refer to models trained with all data. "not-en" refers to model trained without any English data. "only-(vro/sm)" refers to a model trained only on that specific language group data. *prev_best* refers to best results by Tars et al. (2021). **bold** - best BLEU score for a language pair.

process is also one of our development directions and we are planning to subsequently deploy the model to the web to be used freely.

8 Conclusion

Our results show that large pre-trained multilingual translation models significantly improve translation quality for low-resource languages. The more information the model previously has and the more parameters it has, the better is the chance to leverage it during the transfer to low-resource languages. We achieved state-of-the-art results for Võro and Sami translation directions included in our work and achieved comparable results in Livonian to previous fine-tuned model results.

In our analysis of transfer learning between smaller language groups, we found that there is still enough variation between Finno-Ugric languages, which might disturb each other if combined during transfer. Smaller groups like Võro, Livonian and Sami resulted in better translation quality for their respective translation directions.

We created a new training and benchmark dataset, openly shared online to help advance further research in this field.

References

Aharoni, R., Johnson, M., Firat, O. (2019). Massively multilingual neural machine translation, *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 3874–3884.

Aulamo, M., Virpioja, S., Tiedemann, J. (2020). OpusFilter: A configurable parallel corpus filtering toolbox, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Online, pp. 150–156.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V. et al. (2021). Beyond English-Centric Multilingual Machine Translation, *Journal of Machine Learning Research* **22**(107), 1–48.

Gu, J., Hassan, H., Devlin, J., Li, V. O. (2018). Universal neural machine translation for extremely low resource languages, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 344–354.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation, *Transactions of the Association for Computational Linguistics* **5**, 339–351.

Kocmi, T., Bojar, O. (2018). Trivial transfer learning for low-resource neural machine translation, *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, Brussels, Belgium, pp. 244–252.

Kudo, T., Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Brussels, Belgium, pp. 66–71.

Rikters, M., Pinnis, M., Krišlauks, R. (2018). Training and adapting multilingual NMT for less-resourced and morphologically rich languages, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan.

Rikters, M., Tomingas, M., Tuisk, T., Ernštreits, V., Fishel, M. (2022). Machine translation for Livonian: Catering to 20 speakers, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Dublin, Ireland, pp. 508–514.

Sennrich, R., Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 211–221.

Tars, M., Tättar, A., Fišel, M. (2021). Extremely low-resource machine translation for closely related languages, *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), pp. 41–52.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, pp. 2214–2218.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., Polosukhin, I. (2017). Attention is all you need, *in* Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds), *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc.

Zhang, B., Williams, P., Titov, I., Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 1628–1639.

Received August 19, 2022 , accepted August 27, 2022