Teaching Unseen Low-resource Languages to Large Translation Models

Maali Tars, Taido Purason, Andre Tättar
TartuNLP, University of Tartu
maali.tars@ut.ee, taido.purason@ut.ee, andre.tattar@ut.ee

Abstract

In recent years, large multilingual pre-trained neural machine translation model research has grown and it is common for these models to be publicly available for usage and fine-tuning. Low-resource languages benefit from the pre-trained models, because of knowledge transfer from high- to medium-resource languages. The recently available M2M-100 model is our starting point for cross-lingual transfer learning to Finno-Ugric languages, like Livonian. We participate in the WMT22 General Machine Translation task, where we focus on the English-Livonian language pair. We leverage data from other Finno-Ugric languages and through that, we achieve high scores for English-Livonian translation directions. Overall, instead of training a model from scratch, we use transfer learning and back-translation as the main methods and fine-tune a publicly available pre-trained model. This in turn reduces the cost and duration of training high-quality multilingual neural machine translation models.

1 Introduction

We participate in the WMT 2022 General Machine Translation shared task where we submit a system for English-Livonian and Livonian-English translation directions. Our system is trained in the unconstrained setting utilizing data from other languages that are all in a way related to Livonian.

Recently, the development of large multilingual models has been increasing (Johnson et al., 2017; Gu et al., 2018; Fan et al., 2021; NLLB Team et al., 2022) and thus there are multiple pre-trained multilingual models available for further fine-tuning to a specific task. Fine-tuning these models on in-domain data saves time and computational costs by not having to train a multilingual model from scratch. We utilize the M2M-100 multilingual pre-trained neural machine translation (NMT) model (Fan et al., 2021) and do cross-lingual transfer learning to low-resource language pairs from the Finno-Ugric language family, including the Livonian language. We further improve our model with two back-translation iterations and a final fine-tuning on languages that have available original parallel data paired with Livonian.

The languages we use to support the English (en)-Livonian (liv) directions are from the Finno-Ugric language family or geographically close to that family of languages: Finnish (fi), Estonian (et), Latvian (lv), Norwegian (no), Võro (vro), North Sami (sme), South Sami (sma), Inari Sami (smn), Skolt Sami (sms), Lule Sami (smj).

The structure of the article consists of giving insight into the related work in the field of low-resource NMT and from the Finno-Ugric language family perspective in Section 2, the description of data in Section 3, the overview of our system architecture and training methods in Section 4, description of experiments in Section 5 and the results in Section 6.

2 Related work

2.1 Low-resource setting

There have been a lot of efforts in trying to achieve high-quality translation for low-resource languages in order for them to catch up with high- and medium-resource languages. The main benefits seem to come from performing transfer learning to low-resource languages with previous knowledge acquired from a high-resource language (Gu et al., 2018).

Another aspect is data augmentation. Commonly, low-resource languages have a lot more monolingual data available than parallel data, which enables producing synthetic parallel samples that have been shown to improve the accuracy of translation (Sennrich and Zhang, 2019).

For the Finno-Ugric languages, in Rikters et al. (2018), they note that in efforts of achieving better translation quality for Estonian, training a multi-

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 375–380
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

lingual model gets the best result. It usually helps even more if the high- or medium-resourced languages in the mix during training are closely related to the low-resource languages as shown in Tars et al. (2021). In Kocmi and Bojar (2018), the authors proved transfer learning to be very beneficial for languages with low amounts of parallel resources. However, in some cases, they saw more improvements when the high-resource language was not related to the low-resource language.

2.2 M2M-100

M2M-100 is a massively multilingual pre-trained machine translation model featuring many-to-many translations between 100 languages (Fan et al., 2021). It was trained on 7.5 billion parallel sentence pairs which, unlike datasets for many previous approaches, were chosen to make the dataset non-English-Centric. Fan et al. (2021) were able to compose the non-English-Centric training dataset through the use of bitext mining and back-translation. The improvement of M2M-100 over previous models is especially visible in non-English directions and low-resource languages. The vast amount of training data, many supported languages, and promising results reported by Fan et al. (2021) give us reason to believe that M2M-100 would be also a good starting point for training a Finno-Ugric system.

3 Data

3.1 Additional languages

We did not limit ourselves to only English-Livonian training data, because the amount of parallel data for that language pair seemed too scarce to train a quality machine translation system. Instead, we decided to leverage our previous research into Finno-Ugric languages (Tars et al., 2021) and include the language pairs that are closely related to Livonian grammatically as well as geographically.

We added four languages that were high- or medium-resource: Estonian, Finnish, Latvian, Norwegian. The aim of including these languages was for them to aid the low-resource Finno-Ugric languages in the training process. The low-resource languages that we included were: Võro, North Sami, South Sami, Inari Sami, Skolt Sami, Lule Sami.

As Livonian has historically been spoken mainly in the areas that are nowadays Latvia, its language has shaped Livonian noticeably, even though Latvian itself is part of the separate Baltic branch of languages. Multiple low-resource languages that we also included are Sami languages, which are mainly spoken in the areas of Norway, Sweden and Finland. Most of the parallel data available for Sami languages is paired with either Finnish or Norwegian. Norwegian is not part of the Finno-Ugric language family, but as was the case for Latvian, it is spoken in the same area as some of the Sami languages and has influenced them over time, for example sharing some orthographic symbols in the vocabulary.

3.2 Pre-processing and filtering

The data not provided by the shared task was collected from various openly available sources, such as META-SHARE[1] and translation memory compiled by the Arctic University of Norway[2]. Further details about the data sources are described in Tars et al. (2021). We compiled all of the filtered parallel data and the monolingual data and publish it on our HuggingFace page [3].

Following the collection phase, we applied multiple pre-processing and filtering heuristics to the parallel data, as well as deduplicated the whole dataset. We normalized punctuation and detokenized the data with the help of Moses scripts, however, we modified the normalization script for it to be more applicable to Finno-Ugric languages[4]. Detokenization language code defaulted to English if the script did not recognize the language code of a low-resource language. Filtering and whitespace normalization was done with the OpusFilter tool (Aulamo et al., 2020). We provide a list of filters used:

- maximum segment length: 1000 characters or 400 words
- maximum word length: 50 characters
- source and target segment length difference: max 3 times
- ratio of numeric characters in segment: 0.5 or less

[1] https://doi.org/10.15155/1-00-0000-0000-0000-001A0L

[2] https://giellalt.uit.no/tm/TranslationMemory.html

[3] https://huggingface.co/datasets/tartuNLP/finno-ugric-train

[4] https://github.com/Project-MTee/model_training/blob/main/normalization.py

lang-pair	et-vro	fi-sme	fi-sma	fi-smn	fi-sms	no-sma	no-sme	no-smj	sme-sma	sme-smj	sme-smn	en-liv	et-liv	lv-liv
filtered	29 775	62 837	2766	9459	2708	15 702	195 970	11 627	19 963	14 985	894	280	12 887	10 763

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

language	vro	sma	sme	sms	smn	smj	liv
nr of segments	162 807	55 088	33 964	76 685	122 916	128 180	40 329

Table 2: Monolingual data numbers.

- ratio of alphabetic characters in latin alphabet: 1

- ratio of alphabetic characters in segment: 0.75 or more

- ratio of similar numerals between segments, with zeros removed: 0.5 or more

Some of the values are default from OpusFilter, but others had to be tuned to filter out the noisy training samples that were left undetected with the default parameters. The data numbers of all the parallel data for all of the translation directions left after filtering can be seen in Table 1. Additionally, we sampled 20 000 segments from corpora available in OPUS (Tiedemann, 2012) for each language pair between high- to medium-resource languages (et, en, lv, no, fi).

3.3 Monolingual data

We also gathered monolingual data for all the languages involved. The monolingual data for high- and medium-resourced languages was acquired from previously available WMT sources. For the low-resource languages, the data was scraped from various files from the web, that were collected either by the Arctic University of Norway or ourselves.

We sampled 500 000 random segments for all of the high- to medium-resource languages from publicly available data (et, en, lv, no, fi). The amounts of monolingual data for low-resource languages can be seen in Table 2. No filtering was done to the monolingual data, but the data segments all went through the same detokenization and normalization scripts that were applied to parallel data. After the back-translation iterations explained in Section 4.2, the synthetic parallel samples were also not filtered.

For the English-Livonian directions, the only parallel and monolingual data used was the data provided by the WMT.

3.4 Evaluation benchmarks

In order to evaluate the multiple translation directions we had other than English-Livonian, we created new test sets[5] for them, because there are no publicly available benchmarks for translation directions like Finnish-Inari Sami, for example. The test sets are composed of held-out data from the parallel datasets. For all the language pairs containing at least one of the low-resource languages, we extracted 500 sentences for evaluation and 200 sentences for validation.

4 System overview

4.1 M2M-100 settings

Our final system builds on the large pre-trained multilingual neural machine translation model M2M-100. Livonian along with other low-resource Finno-Ugric languages were not part of the training process of M2M-100. We use the HuggingFace implementation of M2M-100[6]. Fine-tuning this model for previously unseen languages requires introducing new symbols to the vocabulary and increasing the embedding matrix. We created scripts[7] that allow expanding the embedding matrix of a pre-trained model and thus make it possible to do cross-lingual transfer learning.

4.2 Stages of training

This section describes the training of our final system. The first stage of transfer learning used all of the original Finno-Ugric parallel data that we had. We decided to go with the M2M-100's 1.2 billion parameter model (1.2B) as our starting point because our previous experiments showed that it improves more than the smaller, 418 million parameter model (418M) on the data that we have (Tars et al., 2022).

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

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

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

After that, we performed the first iteration of back-translation with all of the monolingual data. Combining the original parallel data and the synthetic data, we fine-tuned the M2M-100 1.2B model again and performed the second iteration of back-translation with the newly fine-tuned model. The monolingual data stayed the same.

For the next step we went back to do transfer learning from the beginning on the 1.2B model, but this time the data we used consisted of the original parallel data and the data produced in the second iteration of back-translation, leaving the data from the first iteration out. Finally, we fine-tuned the model on original parallel data for language pairs between `en-liv-et-lv`.

5 Experiments

5.1 Experimental settings

All our systems, including the final system, were trained on one Tesla A100 GPU with 40GB vRAM. Our experiments were done on two versions of the M2M-100 model: 418M model and 1.2B model. The learning rate was initialized with the default value from HuggingFace code. Batch size was 12 with gradient accumulation steps set to 8.

5.2 Different experiments

The size of the model was one aspect of experimentation that we looked into. As smaller models are easier and quicker to fine-tune and deploy, comparing the 418M and 1.2B models seemed necessary. 1.2B model has more parameters, but the intuition was that maybe the 418M model is also big enough for this specific dataset, because it is relatively small.

The main approach to enhance `en-liv` results was leveraging information from other Finno-Ugric languages. We trained models on all the Finno-Ugric language data described, as well as dividing the languages into even smaller groupings, as described in Tars et al. (2022). Subsequently, we performed additional experiments to see whether the added languages really help the Livonian language.

We repeated the stages of training described in Section 4.2 but with different-sized models and with a smaller dataset, consisting only of languages paired with Livonian.

	COMET-A ↓	ChrF-all
`en-liv`	-36.8	39.2
`liv-en`	-5.8	53.5

Table 3: Automatic metric results of our primary system on WMT22 test set.

6 Results

6.1 Automatic metrics

According to the automatic metric results, our system performs the best in the Livonian-English translation direction and achieves second place in the English-Livonian direction. The metrics that were used were COMET and ChrF. The results of the automatic evaluation can be seen in Table 3.

During the development period, we measured most of our additional experiments on BLEU. The results of those experiments compared to the earlier results for English-Livonian translation directions can be seen in Table 4. For further understanding of where the gain in performance happened, we describe the results of intermediate models that were trained before arriving at the final system.

Firstly, we can observe that `en-liv` results are about half of the `liv-en` results and that the BLEU score improvements come from different techniques for either of the translation directions. For `en-liv`, the main source of improvement is the last stage of fine-tuning the model on the original parallel `en-et-lv-liv` data. For `liv-en` however, the biggest gain happens with back-translation. This could be explained by the amount of monolingual data, as Livonian had only about 40 000 segments but for English, we sampled 500 000 segments.

Another aspect we can point out is the relatively small difference between the smaller (418M) and the larger (1.2B) model results. The 1.2B model is better at every stage as expected, but considering how much more computational cost and deployment resources the larger model requires, the trade-off in quality might be tolerable.

Lastly, compared to the previous best results reported by Rikters et al. (2022), our models surpass those results by about 4 BLEU for `en-liv` and 12 BLEU for `liv-en`.

6.2 Results for other language pairs

Additionally, we report results on our held-out test set described in Section 3.4 for low-resource language pairs that were a part of our final system development. The results can be seen in Table

	en-liv	liv-en
1.2B (baseline)	10.15	18.92
+ bt1	11.24	28.67
+ bt2	12.16	29.37
+ tuned on liv	**15.19**	**31.06**
+ bt1 only-liv	10.66	27.88
+ bt2 only-liv	11.21	29.85
+ tuned on liv	11.56	30.33
Rikters et al., 2022	*11.03*	*19.01*

	en-liv	liv-en
418M (baseline)	10.29	15.78
+ bt1	11.25	27.52
+ bt2	10.62	27.38
+ tuned on liv	**12.83**	27.23
+ bt1 only-liv	11.39	27.74
+ bt2 only-liv	11.63	28.81
+ tuned on liv	11.53	**29.27**
	11.03	*19.01*

Table 4: Experiment results on BLEU. “1.2B” and “418M” refer to models trained with all original parallel data. “bt1” is trained on parallel + first back-translation iteration data, “bt2” on parallel + second back-translation iteration data. “only-liv” - only data between et-en-lv-liv languages was used for training. “tuned on liv” refers to the “bt2” model that was tuned on et-en-lv-liv original parallel data. Last row represents previously best results for en-liv-en by Rikters et al. (2022).

5 and Table 6. The language pairs were evaluated on the final system and although the final system was chosen on en-liv-en validation data, we see good overall results for other low-resource language pairs as well. However, the results in Table 5 are significantly lower than the results reported in Tars et al. (2022) on the same test data. This is probably caused by the fact that as the last training stage, the final system was fine-tuned only on et-en-lv-liv original parallel data.

For et-liv-et and lv-liv-lv directions, however, we report new state-of-the-art results on the test data that was also used in Rikters et al. (2022).

7 Conclusion

Large pre-trained multilingual neural machine translation models prove to be beneficial to low-resource Finno-Ugric languages, such as Livonian. We placed in the top 2 for the English-Livonian language pair in the WMT22 General Machine Translation shared task. Training in an unconstrained setting gets reasonable and good-quality results, especially when using languages close to Livonian to help achieve a better translation quality. In the future, we plan to test additional and more recent pre-trained multilingual models as a starting point for cross-lingual transfer learning and add more low-resource Finno-Ugric languages into the dataset.

Limitations

The 1.2B M2M-100 model has a lot of parameters which makes deploying this model very costly and difficult because it needs a lot of memory and is computationally unfeasible. In turn, it also makes the training somewhat slower in terms of loading the model parameters and updating them. We are working on trying to reduce the vocabulary and number of parameters, by removing parts of the vocabulary not necessary for Finno-Ugric languages. Another thing we left out of the process was filtering monolingual and synthetic data, which might be a useful addition to the pre-processing pipeline.

References

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

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

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

	WMT22 sys	*previous*
`fi-sma-fi`	12.31	***29.04***
`fi-sme-fi`	36.88	***46.56***
`fi-smn-fi`	53.73	***64.37***
`fi-sms-fi`	36.14	***47.61***
`no-sma-no`	40.59	***50.16***
`no-sme-no`	33.89	***40.61***
`no-smj-no`	32.31	***46.22***
`sme-sma-sme`	26.16	***40.37***
`sme-smj-sme`	22.32	***40.24***
`sme-smn-sme`	31.73	***33.88***
`et-vro-et`	34.76	***37.08***

Table 5: BLEU scores for low-resource language pairs included in the final system. *previous* signifies the previous best results for these language pairs reported in Tars et al. (2022).

	WMT22 sys	*previous*
`et-liv`	**18.31**	*16.49*
`liv-et`	**24.00**	*23.05*
`lv-liv`	**19.16**	*17.65*
`liv-lv`	**26.33**	*25.24*

Table 6: BLEU scores for low-resource language pairs included in the final system. *previous* signifies the previous best results for these language pairs reported in Rikters et al. (2022).

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, Brussels, Belgium. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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

Matīss Rikters, Marili Tomingas, Tuuli Tuisk, Valts Ernštreits, and Mark Fishel. 2022. Machine translation for Livonian: Catering to 20 speakers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 508–514, Dublin, Ireland. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

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

Maali Tars, Andre Tättar, and Mark Fišel. 2022. Cross-lingual transfer from large multilingual translation models to unseen under-resourced languages. *Baltic Journal of Modern Computing*, 10.3:435–446.

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