Language Model Priors and Data Augmentation Strategies for Low-resource Machine Translation: A Case Study Using Finnish to Northern Sámi

Jonne Sälevä and **Constantine Lignos**
Michtom School of Computer Science
Brandeis University
{jonnesaleva,lignos}@brandeis.edu

Abstract

We investigate ways of using monolingual data in both the source and target languages for improving low-resource machine translation. As a case study, we experiment with translation from Finnish to Northern Sámi. Our experiments show that while conventional backtranslation remains a strong contender, using synthetic target-side data when training backtranslation models can be helpful as well. We also show that monolingual data can be used to train a language model which can act as a regularizer without any augmentation of parallel data.

1 Introduction

This paper focuses on improving machine translation (MT) between the Northern Sámi (ISO 639-3: `sme`) and Finnish (`fin`) languages. With approximately 25,000 speakers, Northern Sámi has the most speakers of the Sámi languages, which are spoken by the Sámi people. The Sámi are the Indigenous peoples of Sápmi, which spans modern-day Norway, Sweden, Finland, and Russia.[1]

This paper provides a contribution towards improving Northern Sámi-Finnish machine translation, with the hope that these findings may be generalized to other languages in subsequent work. We advance the state of the art for Northern Sámi-Finnish MT established by Aulamo et al. (2021) using a combination of techniques.

The contributions of this paper are as follows. First, we demonstrate that current LLM-based MT models are insufficient for translating this language pair—and likely other pairs including less-resourced/minoritized languages—by evaluating how well the recent MADLAD-400 model (Kudugunta et al., 2023) performs on our datasets. Second, we demonstrate that the parallel data-only baseline of Aulamo et al. (2021) was undertrained, over-emphasizing the relative improvement of their backtranslation models over it. We find that sufficient training improves baseline model BLEU by nearly 8 points. Third, we experiment with a wide range of configurations involving forward and backtranslation, finding that good model performance is associated with large data size and including rule-based MT system outputs as backtranslation data in addition to neural MT system outputs. Finally, we experiment with combining models trained on noisy target backtranslations with existing techniques such language model priors. All code and data (forward and backtranslation outputs) can be found at `https://github.com/j0ma/sami-translation`.

[1]Further reading: The Sámi Language Crisis, Language Policy and the Sámi Languages: An Investigation Using the Comparative Perspective, and Indigenous Peoples in Sápmi.

2 Related work

Northern Sámi language technology Much of the Northern Sámi language technology research to date has been performed at the Giellatekno Institute at the Arctic University of Norway. Although a large part of the research focuses on building tools like keyboards and dictionaries (Nørstebø Moshagen et al., 2022), recent work using their corpora has also focused on MT (Aulamo et al., 2021). There has also been work on integrating Northern Sámi with existing multilingual MT models such as M2M-100 (Tars et al., 2022a; Yankovskaya et al., 2023; Tars et al., 2022b).

Backtranslation and forward translation Back- and forward translation also have a long history in neural machine translation research. Backtranslation was introduced in the NMT context by Sennrich et al. (2016) where the authors observe 2-3 BLEU increases on English-German and English-Turkish translation tasks. Forward translation was also introduced around the same time by Zhang and Zong (2016) who use it as both a standalone technique and in a multi-task framework.

Findings of the Association for Computational Linguistics: ACL 2024, pages 12949–12956

	UiT Test		YLE Test	
Direction	BLEU	CHRF	BLEU	CHRF
Sámi → Finnish	7.99	30.75	10.62	30.96
Finnish → Sámi	1.92	16.92	3.66	22.30

Table 1: Results using `MADLAD400-3b-mt`.

Regularization via language models Language models also have a long history of being used as regularizers in the NMT literature. However, many approaches seem to focus on approaches like reranking at test time (e.g. Luong and Popescu-Belis, 2016), rather than using the LM at train time. Baziotis et al. (2020) introduce a technique to do this via a KL-divergence objective and report performance gains of several BLEU points on English-German and English-Turkish translation tasks.

Massively multilingual language models Recently, Kudugunta et al. (2023) introduced MADLAD-400, a massively multilingual encoder-decoder language model (MMLM) with support for both Finnish and Northern Sámi. To gauge whether our translation task has already been solved by developments in MMLMs, we evaluated the performance of the 3B parameter MT variant of MADLAD-400, `madlad400-3b-mt`, using test sets we describe later. Based on the results given in Table 1, the translation quality is unusably poor, especially when translating into Northern Sámi.[2]

3 Methodology

We expand upon the experiments of Aulamo et al. (2021) in several ways: using more powerful model architectures for backtranslation (Transformer instead of RNN), more elaborate data augmentation including synthetic target-side data, and Transformer-based autoregressive language model priors.

3.1 Improved data augmentation

Model architecture Unlike the RNN and rule-based translation architectures that Aulamo et al. (2021) use for backtranslation, we rely exclusively on Transformer models when extending their work.

Synthetic target-side data for backtranslation In addition to training a reverse Northern Sámi-Finnish Transformer using the 25k parallel sentences, we further augment the training data of the reverse model Northern Sámi-to-Finnish with synthetic backtranslations produced by the RBMT model of Aulamo et al. (2021). That is, we train using a dataset which contains synthetic model outputs as targets in the training set. After training the models on the parallel and RBMT-augmented data sets, we use them to translate the monolingual Northern Sámi data into Finnish which yields two sets of backtranslations, one from each model. A diagram of our approach can be seen in Figure 1 in the Appendix.

Forward translation Since we have access to monolingual data in Finnish, we use forward translation (Bogoychev and Sennrich, 2020) to further augment the training data. We use each Finnish-Northern Sámi model to translate another ∼950k sentences from the mC4 corpus (Xue et al., 2021) into Northern Sámi. This approach is very similar to the "self-training" setup of Zhang and Zong (2016).

Tagged backtranslation Applied together, the backtranslation and forward translation procedures yield datasets as large as 1.9M sentences which is significantly larger than the original dataset of 25,106 parallel sentences. To avoid "crowding out" the authentic parallel data we also experiment with prepending `<BT>` as the first token to the the synthetic sentences.

3.2 Language model prior

We experiment with alternatives to data augmentation using an autoregressive Transformer language model, $p_{\text{LM}}(y_t|y_{<t})$, as a weakly informative prior on $p_{\text{NMT}}(y_t|x, y_{<t})$, as introduced by Baziotis et al. (2020). We train a language model on the Northern Sámi monolingual data and augment the token-level NMT log-likelihood loss by adding a KL-divergence regularization term, yielding $J(\theta) = -\log P_{\text{NMT}}(y|x) + \lambda D_{\text{KL}}(p_{\text{LM}}(y)|p_{\text{NMT}}(y|x))$. The LM only serves as a regularizing prior and remains frozen through the NMT training; gradient updates are only applied to the NMT model. The regularization term encourages the distribution learned by the NMT model to be similar to that of the language model.

[2]While it may be possible to further finetune massively multilingual models to improve performance in less-resourced languages, in preliminary experiments our performance remained under 10 BLEU when using quantized LoRA finetuning (Dettmers et al., 2023).

Corpus	Sentences
Full data sets (Aulamo et al., 2021)	
Parallel data	25,106
RNN	470,098
RBMT	487,862
RBMT + RNN	932,803
Cleaned data sets (Aulamo et al., 2021)	
RBMT + clean	378,567
RNN + clean	422,608
RBMT-clean + RNN-clean	610,093
RNN-clean + RBMT-all	885,313
Our data augmentations	
Transformer-BT (combo)	950,712
Baseline + fwd. translation	938,071
Transformer-BT (combo) + fwd. translation	1,949,992
Test and validation sets	
UiT valid/test sets (each)	2,000
YLE test set	151

Table 2: Train, validation and test data sizes (number of sentences) across our experiments.

4 Experimental results

4.1 Datasets and model

Datasets We use the parallel data and the backtranslation outputs produced by the RNN and RBMT models of Aulamo et al. (2021). For forward translation, we use a ∼950k sentence sample of the Finnish subset of the mC4 corpus (Xue et al., 2021). We evaluate on two test sets: UiT (in-domain, web) and YLE (smaller, out-of-domain, news). Our dataset sizes can be seen in Table 2.

All data are tokenized into subwords using the unigram language model subword tokenization method as implemented in the SentencePiece library (Kudo and Richardson, 2018). We use a separate models for source and target languages. All segmentation models learn vocabularies of 8,000 units except for the language model-augmented experiments (5,000 units).

Model Our NMT architecture is an encoder-decoder Transformer similar to `transformer-base` by Vaswani et al. (2017), i.e. 6 layers in both the encoder and the decoder, 512-dimensional embeddings and 2048-dimensional feedforward layers. All NMT models are trained using the label smoothed cross-entropy objective with label smoothing probability set to 0.1. For the language model-augmented experiments we use a similarly sized decoder-only Transformer LM with 512-dimensional embeddings and 2048-dimensional hidden layers. Further details are provided in Appendix A.1.

4.2 Experiment 1: Transformer-based backtranslation

We sought to replicate the original results by Aulamo et al. (2021) and trained Transformer-based translation models using their data. In addition to replicating training on backtranslation data generated by their RNN and RBMT-based models, we also trained Transformer-based models in the Finnish-to-Northern Sámi direction on our own backtranslation data. These backtranslations were produced by Northern Sámi-to-Finnish models trained on either the parallel data or the combination of parallel data and backtranslations produced by the original RBMT model.

Our experimental results can be seen in Tables 3 and 4. From Table 3 it is clear that the baseline model evaluated by Aulamo et al. (2021) was undertrained. Using the same data, our baseline model achieves an almost 8 BLEU absolute improvement over the previous baseline of 18.90 BLEU, yielding 26.86 BLEU. With the exception of the baseline and the Transformer model trained using RNN-generated backtranslation data, the rest of our replicated scores are less than those reported by Aulamo et al. (2021). We attribute this discrepancy to differences between fairseq (Ott et al., 2019) and MariaNMT (Junczys-Dowmunt et al., 2018), as we used the same model architecture.

Our best model is the Transformer-BT model trained on two sets of backtranslation outputs which achieves 41.75 BLEU, a gain of 1.65 BLEU over the best 40.1 BLEU performance reported by (Aulamo et al., 2021). Notably, not all Transformer-based BT models outperform RNN and RBMT-based backtranslation. The first two Transformers, trained on 25k and 487k sentences respectively, seem to underperform whereas the third Transformer trained on ∼900k sentences achieves 41.75 BLEU. Based on Table 4, on the out-of-domain YLE test set, our replicated models outperform the scores reported by Aulamo et al. (2021) on all configurations except "RNN + clean" and "RBMT + clean". Our best out-of-domain score is attained by Transformer-BT (RBMT) which achieves 16.34 BLEU on the YLE test set.

4.3 Experiment 2: Beyond backtranslation

Next, we investigated to what extent other techniques can be used in conjunction with or in lieu of

	Finnish–Northern Sámi					Northern Sámi–Finnish			
	UiT Test			UiT Valid		UiT Test		UiT Valid	
Reverse model	Aulamo	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF
Baseline	18.9	26.86	58.60	25.46	57.99	27.30	59.51	25.40	58.94
RNN	32.9	33.51	62.73	31.85	61.77	24.34	56.81	24.25	56.83
RBMT	37.0	35.85	66.86	34.26	65.93	26.86	61.08	27.10	61.82
RBMT + RNN	38.8	36.76	62.73	34.95	64.85	23.56	58.23	22.82	61.18
RNN + clean	34.0	33.23	62.69	31.93	62.04	24.48	56.78	24.76	56.96
RBMT + clean	36.3	34.46	66.26	33.34	65.70	26.79	60.87	26.65	58.40
RNN-clean + RBMT-clean	38.9	36.88	66.24	35.80	65.67	22.21	59.38	21.84	57.62
RNN-clean + RBMT	40.1	37.48	66.65	35.54	65.58	25.51	59.38	24.43	59.55
Transformer-BT (baseline)	-	32.91	64.21	31.60	63.51	25.30	59.55	25.65	59.97
Transformer-BT (RBMT)	-	37.22	67.89	36.46	67.58	**28.21**	**61.84**	**28.01**	**62.41**
Transformer-BT (combined)	-	**41.75**	**70.12**	**40.37**	**69.49**	27.29	61.05	27.66	61.70

Table 3: UiT in-domain results using Transformer-based models trained on backtranslated data produced by various reverse models. The highest values are bolded for each translation direction, dataset, and metric. The column labeled *Aulamo* contains BLEU scores reported by Aulamo et al. (2021).

	Finnish–N. Sámi			N. Sámi–Finnish	
Reverse model	Aulamo	BLEU	CHRF	BLEU	CHRF
Baseline	4.3	8.43	35.67	5.74	33.88
RNN	9.2	9.84	35.80	7.54	32.23
RBMT	14.4	15.92	45.65	10.79	39.21
RNN + RBMT	10.9	14.76	41.12	9.30	38.70
RNN + clean	9.8	9.64	35.75	7.48	33.18
RBMT + clean	15.5	14.90	45.74	**11.03**	**39.96**
RNN-clean + RBMT-clean	11.3	14.79	42.13	10.11	39.21
RNN-clean + RBMT	10.8	13.53	41.54	10.14	38.75
Transformer-BT (baseline)	-	10.92	40.34	8.30	36.36
Transformer-BT (RBMT)	-	**16.34**	**47.46**	10.00	38.46
Transformer-BT (combined)	-	14.96	45.93	10.20	39.08

Table 4: YLE out-of-domain test set results. The highest values are bolded for each translation direction and metric. The column labeled *Aulamo* contains BLEU scores reported by Aulamo et al. (2021).

backtranslation (Table 5). Since our goal is to improve Finnish-Northern Sámi translation, we only evaluate in that direction in this experiment.

Tagging First, we took our best performing Finnish-Sámi model and prepended a <BT> tag to all of the backtranslated sentences. This proved beneficial in some scenarios but not others: the new tagged Transformer-BT model achieved 41.71 BLEU on the UiT test set, whereas the non-tagged version achieved 41.75 BLEU. In terms of CHRF, the model trained on tagged data achieves 70.33 CHRF, a slight improvement over 70.12 CHRF without tagging. In terms of out-of-domain performance, we note that the Transformer-BT model trained on tagged data outperforms its non-tagged counterpart by 4.24 BLEU (19.20 vs 14.96) and 2.67 CHRF (48.60 vs. 45.93).

Forward translation (FT) We applied the model to ∼950k sentences of Finnish monolingual data, creating another set of synthetic parallel data. We created two augmented datasets: (a) original parallel data combined with the forward translated sentences. and (b) the ∼900k sentences used to train the best-performing Transformer model combined with the forward translated sentences.

The results using this approach were mixed. With our best Transformer model, FT decreased performance from 41.75 BLEU to 38.84 BLEU (with <BT> tagging) and 38.96 BLEU (no tagging) on the UiT test set. Augmenting the baseline with FT data yielded 25.93 BLEU on the UiT test set compared to the original 26.86 BLEU. On the out-of-domain YLE test set, the baseline model benefitted slightly from forward translation and achieved 9.01 BLEU, a modest increase over the

	UiT Test		UiT Valid		YLE Test	
Reverse model	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF
Baseline	26.86	58.60	25.46	57.99	8.43	35.67
+ forward translation	25.93	59.52	25.10	59.03	**9.01**	36.99
+ language model	**31.26**	**62.99**	**30.18**	**62.40**	8.60	**37.41**
Transformer-BT (combined)	**41.75**	70.12	**40.37**	**69.49**	14.96	45.93
+ tagging	41.71	**70.33**	39.73	69.44	**19.20**	**48.60**
+ forward translation	38.96	69.16	37.44	68.36	17.91	**48.60**
+ forward translation + tagging	38.84	69.10	37.55	68.57	17.36	48.19
+ language model	33.30	65.69	31.45	64.77	14.51	45.52

Table 5: Effect of various techniques on baseline model and the best-performing Transformer-BT model.

original 8.43 BLEU score. The best Transformer-BT model experienced a slightly larger boost and outperformed the original 14.96 BLEU score by 2.95 BLEU, scoring 17.91 BLEU when using forward translation.

Although the scores were generally lower on the UiT in-domain data, this may be partially due to domain effects and should not be taken as a fully general dismissal of forward translation as a method. The sentences in the UiT data set are largely drawn from political texts whereas the forward translation data is more general in nature. We conjecture that may able to mitigate in-domain overfitting as shown by the out-of-domain results. Whether there is a benefit from forward translation with domain-aligned data remains to be seen.

Language model prior Finally, we investigated whether a Transformer LM could be used as an alternative regularizer compared to the costly process of training forward and reverse models. We took our baseline and best-performing Transformer models (without `<BT>` tagging) and re-trained them from scratch using the same data plus an LM prior which we trained on the ∼460k Sami monolingual data otherwise used for backtranslation. While this approach improved over the baseline on both UiT (31.26 BLEU vs. 26.86 BLEU) and YLE (8.60 BLEU vs. 8.43 BLEU), the performance was still lower than forward translation (8.60 BLEU vs 9.01 BLEU).

Using the best Transformer model, the language model approach significantly decreased BLEU on UiT (33.80 BLEU vs. 41.75 BLEU) and slightly on YLE (14.51 BLEU vs. 14.96 BLEU). This suggests that the backtranslated data may already be providing enough regularization. We also hypothesize that with proper hyperparameter tuning the LM prior and forward translation could be beneficial when used together. However, in preliminary experiments we observed this setup to only hurt performance. Whether this is due to issues such as domain mismatch or these methods *per se* is unclear at this point.

5 Conclusion

We have shown several ways to augment extremely small parallel datasets with synthetic data to enable performance improvements. In the absence of publicly-available large monolingual data sets, language models may provide another avenue for regularization, potentially even using off-the-shelf language models.

A key question is whether the approaches described in this work will generalize well to other languages. We were not able to answer this question primarily due to computational resource considerations, as we only had enough compute resources to run our experiments in the language pairs explicitly supported by our funding source.

Acknowledgements

Author JS was supported by a grant by the Lapland Regional Fund of the Finnish Cultural Foundation, awarded with the goal of building modern language technology solutions for Arctic minority languages, including Northern Sámi.

Limitations and risks

The major limitations of our experiments are that we only examine a single language pair and use only the results of a single run for each experiment. A potential risk of our work is that readers might extrapolate the results too much without taking into account that the findings may not generalize beyond the language pair we study. We believe it is likely that the data augmentation strategies we experimented with will be advantageous in many

other settings, but we do not expect that others will find results completely consistent with ours.

Our work also deals with an Indigenous language, which opens the door for potential misuse of NLP models against marginalized peoples. Even when used by those peoples themselves, quality issues with the models (such as mistranslations) can have negative impact due to providing incorrect information. This can be especially problematic with out-of-domain data as our experiments have shown. However, as our models establish a state of the art for the datasets we examine and our limited analysis of a popular multilingual model (MADLAD-400) suggests its performance is much lower than our models, we believe the risks from our model are less than those of existing ones.

Ethical considerations and broader impact

Our research involves work with the Northern Sámi language, a minoritized language spoken by Indigenous persons, and any such research merits scrutiny (Eriksen et al., 2021). We believe the broader impact of our work will be improved translation for these language communities, eventually enabling better communication and scenarios such as Northern Sámi speakers from other countries being able to access information in Finnish relevant to them that was previously inaccessible. All MT models come with inherent risks and potential negative impact, but we believe that by creating improved models, we are not increasing those risks and are making the impact of the models more positive.

References

Mikko Aulamo, Sami Virpioja, Yves Scherrer, and Jörg Tiedemann. 2021. Boosting neural machine translation from Finnish to Northern Sámi with rule-based backtranslation. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 351–356, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online. Association for Computational Linguistics.

Nikolay Bogoychev and Rico Sennrich. 2020. Domain, translationese and noise in synthetic data for neural machine translation.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Heidi Eriksen, Arja Rautio, Rhonda Johnson, Catherine Koepke, and Elizabeth Rink. 2021. Ethical considerations for community-based participatory research with Sami communities in North Finland. *Ambio*, 50:1222–1236.

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.

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

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. 2023. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Ngoc Quang Luong and Andrei Popescu-Belis. 2016. A contextual language model to improve machine translation of pronouns by re-ranking translation hypotheses. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 292–304.

Sjur Nørstebø Moshagen, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen, and Per Langgård. 2022. European language equality: D1. 38: Report on the nordic minority languages.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

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

Maali Tars, Taido Purason, and Andre Tättar. 2022a. Teaching unseen low-resource languages to large translation models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 375–380, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maali Tars, Andre Tattar, and Mark Fishel. 2022b. Cross-lingual transfer from large multilingual translation models to unseen under-resourced languages. *Baltic Journal of Modern Computing*, 10(3):435–446.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771, Tórshavn, Faroe Islands. University of Tartu Library.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

A Appendix

A.1 Model details

We ran all of our experiments using the `fairseq` framework (Ott et al., 2019) on a mixture of NVIDIA V100, A5000, and RTX 3090 GPUs. Each training run used a single GPU; we simulate training on multiple GPUs by accumulating gradients for 24 backward passes using the `update_freq` option in fairseq. For the baseline, we lowered this to 8 backward passes to stabilize training dynamics. Each batch contains a maximum of 16-17,000 tokens, depending on how much fits on each GPU. For the baseline, we lowered the batch size to 1,000 tokens to stabilize training and prevent overfitting. Each run took approximately 18 GPU hours.

A figure showing the relationship between the original data, backtranslated data, and the models we trained appears on the next page.

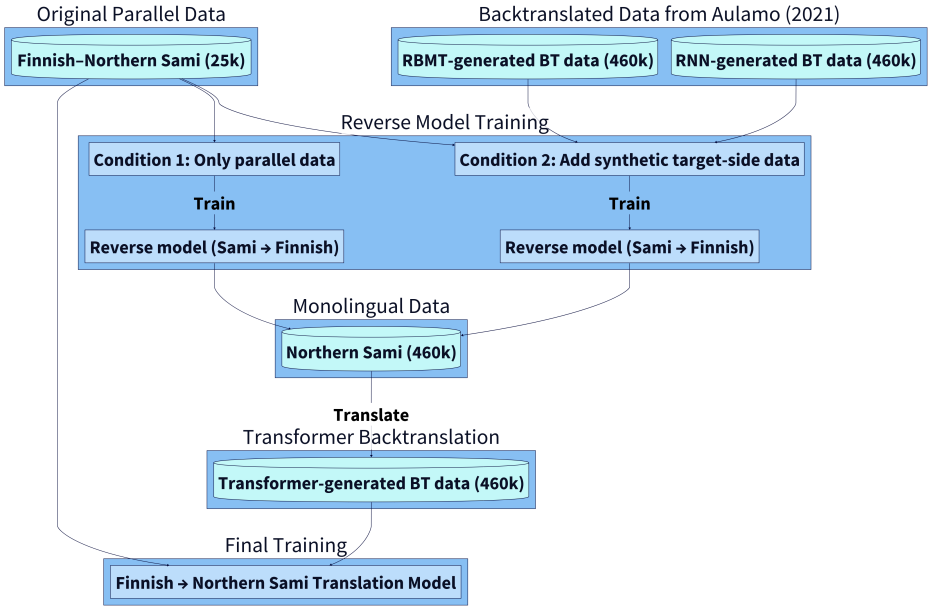

Figure 1: Transformer-based backtranslation setup used in Experiment 1. In addition to the original parallel data, we use RBMT-generated synthetic Finnish data to train Finnish-to-Northern Sámi models. We use them to generate further backtranslations on which we train the three Transformer-BT models in Tables 3 and 4.