Boosting Neural Machine Translation from Finnish to Northern Sámi with Rule-Based Backtranslation

Mikko Aulamo, Sami Virpioja, Yves Scherrer, Jörg Tiedemann
Department of Digital Humanities
University of Helsinki, Helsinki / Finland
{name.surname}@helsinki.fi

Abstract

We consider a low-resource translation task from Finnish into Northern Sámi. Collecting all available parallel data between the languages, we obtain around 30,000 sentence pairs. However, there exists a significantly larger monolingual Northern Sámi corpus, as well as a rule-based machine translation (RBMT) system between the languages. To make the best use of the monolingual data in a neural machine translation (NMT) system, we use the backtranslation approach to create synthetic parallel data from it using both NMT and RBMT systems. Evaluating the results on an in-domain test set and a small out-of-domain set, we find that the RBMT backtranslation outperforms NMT backtranslation clearly for the out-of-domain test set, but also slightly for the in-domain data, for which the NMT backtranslation model provided clearly better BLEU scores than the RBMT. In addition, combining both backtranslated data sets improves the RBMT approach only for the in-domain test set. This suggests that the RBMT system provides general-domain knowledge that cannot be found from the relative small parallel training data.

1 Introduction

Machine translation from and to minority languages is challenging because large parallel corpora are typically hard to obtain. Two strategies have proven most successful to eliminate this bottleneck: using rule-based machine translation (RBMT) systems that do not rely on large data, or training data-driven translation systems with automatically created synthetic data, e.g. backtranslation (Sennrich et al., 2016). In this paper, we combine both strategies in the context of neural machine translation (NMT) from Finnish to Northern Sámi. In particular, we investigate the impact of RBMT in data augmentation in comparison to standard NMT-based backtranslation.

Northern Sámi is a Uralic minority language spoken in Norway, Sweden and Finland. Historically, most of the work on machine translation from and to Sámi languages is based on RBMT (Trosterud and Unhammer, 2012; Antonsen et al., 2017; Pirinen et al., 2017). Data-driven approaches such as NMT are generally more competitive, but require large amounts of training data in the form of parallel translated sentences. For minority languages, finding parallel data sets is usually more difficult than collecting monolingual data, which is also the case for Northern Sámi.

A common way of leveraging monolingual data for NMT is the above mentioned backtranslation strategy, a method where monolingual data of the target language is translated automatically to the source language to create additional parallel training data. In this work, we use two reverse translation models to produce the backtranslations: a neural model trained only on the available parallel data and a rule-based approach. The latter is a system developed for the translation from Northern Sámi to Finnish (Pirinen et al., 2017) within the Apertium framework (Forcada et al., 2011). We also combine both methods to further augment the data. Our experiments demonstrate the positive effects of both strategies and the possibility of obtaining complementary information from different backtranslation engines.

2 Related work

Using backtranslations from different sources as training data has been shown to be beneficial for improving machine translation quality. In addition to proposing training data augmentation methods that do not require reverse translation systems, Burlot and Yvon (2018) compare the effects of using statistical machine translation (SMT) and NMT based backtranslations for English→French and English→German translations. They show that both types of backtranslations improve translation quality, NMT slightly more than SMT. Poncelas et al. (2019) also produce backtranslations with SMT and NMT. They show that the translation quality of a German→English NMT system is improved when including either type of backtranslations in the training data. The greatest improvement is observed when both types of backtranslations are used.

Augmenting training data with RBMT backtranslations has also proven to be useful for boosting translation quality. Dowling et al. (2019) use RBMT backtranslations to improve statistical machine translation performance for Scottish Gaelic→English translations. The authors show that backtranslations can be beneficial even in cases where the translation quality of the MT system used to produce the backtranslations is low. Soto et al. (2019) study the performance of NMT systems trained with augmented training data backtranslated using RBMT, SMT and NMT. They experiment with Basque→Spanish translations and show that the translation performance improves when using each type of augmented training data individually. Soto et al. (2020) also analyze the effects of using augmented training data backtranslated with the three different paradigms. They focus on two language pairs: a low-resource language pair, Basque→Spanish, and a high-resource language pair, German→English. In addition to showing similar results as Soto et al. (2019), they show further improvement in translation performance when all types of augmented training data are combined.

3 Data

The UiT freecorpus[1] contains a Finnish - Northern Sámi (fin-sme) parallel corpus with 110k sentence pairs and a distinct set of 868k monolingual Northern Sámi sentences. The UiT corpora are collected from multiple sources and cover various domains. Both the parallel and the monolingual corpora contain considerable amounts of duplicate lines. In this section, we describe our data cleaning and filtering efforts and the data split. For additional evaluation, we collected a small test set consisting of translated YLE news articles[2].

Data filtering and cleaning is carried out with the OpusFilter toolbox (Aulamo et al., 2020). Our OpusFilter configuration files are available online[3], which helps to replicate the data preprocessing steps. First, we remove duplicate lines from the parallel corpus. This process removes 67.7% of the sentence pairs, leaving us with 35,426 unique sentence pairs. The remaining data set is then cleaned with a set of filters from OpusFilter. Similar filtering setups have been confirmed to improve translation quality (Vázquez et al., 2019; Aulamo et al., 2020). In particular, we remove sentence pairs that satisfy one of the following conditions:

- One or both of the sentences are empty or longer than 100 words,
- The ratio of the sentence lengths in words is greater than 3,
- The sentence pair contains words longer than 40 characters,
- The sentence pair contains HTML elements,
- The sentences have dissimilar numerals based on the "Non-zero numerals score" (Vázquez et al., 2019),
- The sentences have dissimilar punctuation based on the "Terminal punctuation score" (Vázquez et al., 2019),
- The sentence pair contains characters outside of the Latin script,
- The sentences are not recognized to be their correct language by the `langid.py` language identifier (Lui and Baldwin, 2012).

After filtering, 29,106 clean sentence pairs remain in the parallel data set. From this clean set, 2000 pairs are randomly selected to form a validation set and another 2000 pairs to form a test set, leaving 25,106 pairs for training. Note that all subsets are disjoint due to the initial deduplication.

The additional test set consists of two news articles describing Sámi culture in Finland available in both Finnish and Northern Sámi on YLE News. It was extracted from the web and manually aligned to create a clean reference set. This test set

[1] https://giellatekno.uit.no/

[2] https://yle.fi/uutiset/osasto/sapmi/

[3] https://github.com/Helsinki-NLP/Sami-MT

is, however, small (151 sentence pairs) and may not produce completely reliable evaluation scores, but it should still provide additional insights about the quality of the translation models and their ability to generalize to new domains.

The monolingual Northern Sámi data is processed in a similar way as the parallel data above. Duplicate removal discards 35.6% of the total of 867,677 sentences, leaving 559,074 sentences in the data set. For corpus cleaning, we use all filters of those cited above that are applicable to monolingual data, i.e. the sentence length filter, the word length filter, the HTML element filter, the Latin script filter, and the language identification filter. The resulting clean monolingual corpus contains 462,803 sentences.

4 Method

In this section, we compare a baseline fin-sme NMT model trained only with the available parallel data to NMT models trained with additional backtranslated data. The backtranslations are produced by translating the clean monolingual Northern Sámi data to Finnish either with a NMT system trained on the parallel data in the reverse direction (sme-fin), or with the sme-fin RBMT system. This yields three additional synthetic training sets that augment the original parallel training data: one with the NMT backtranslations, one with RBMT translations, and one with both types of backtranslations. Each of them is then used to train a separate NMT model that we can compare to the baseline model, which is trained on the original parallel data only. Note that we do not use any data sampling or weighting scheme to balance original and augmented training data.

All NMT models in our experiments are trained with MarianNMT (Junczys-Dowmunt et al., 2018) version `1.8.33`. The backtranslation model is based on a RNN architecture with GRU cells (Cho et al., 2014) and attention. In our experiments, the RNN architecture slightly outperformed Transformers in the out-of-domain test set for this translation direction. All models using additional backtranslated training sets are trained with both RNNs and Transformers. All RNN models have the same architecture as the backtranslation model. For Transformers, we use the example hyperparameters from MarianNMT [4] which replicate the setup from Vaswani et al. (2017). For subword segmentation, we use the SentencePiece tokenizer (Kudo and Richardson, 2018) with vocabulary size 8000, which has been shown to produce the best results with the data set sizes that we are dealing with (Gowda and May, 2020; Grönroos et al., 2021). We train the models until the cross-entropy of the validation set does not improve for 10 consecutive validation steps.

	UiT	YLE
NMT	19.4	4.5
RBMT	12.3	10.0

Table 1: Reverse translation model (sme-fin) quality in BLEU points evaluated with the UiT test set and the YLE test set.

For the RBMT backtranslations, we use Apertium with the sme-fin model by Pirinen et al. (2017). This system implements a shallow transfer-based translation engine consisting of modules for morphological analysis, disambiguation and generation, modules for lexical translation based on context rules, and a module for syntactic transformation operations.

Table 1 shows the quality of the sme-fin translation models used for backtranslations in BLEU points (Papineni et al., 2002). The NMT model performs much better with UiT test data than with the YLE test data, which shows that the NMT system is strongly adapted to the UiT data, while the RBMT system has similar performance with both test sets.

4.1 Backtranslations

All the 462,803 sentences of the cleaned monolingual data are translated with the sme-fin NMT and RBMT models. As the quality of the source side of the backtranslations is not as important as the quality of the target side (Sennrich et al., 2016), we keep an unfiltered version of both backtranslation data sets. To see the effect of filtering the augmented data set, we apply OpusFilter with a reduced set of filters (recall that the monolingual Northern Sámi data has already been processed): sentence length filter, length ratio filter, word length filter, HTML element filter, nonzero numeral filter and terminal punctuation filter. After filtering and an additional deduplication step, the NMT-produced backtranslations amount to 415,313 sentence pairs and the RBMT-

[4] `https://github.com/marian-nmt/marian-examples/tree/master/transformer`

		Transformer		RNN	
	Training data	UiT	YLE	UiT	YLE
Baseline	25,106	18.9	4.3	18.5	5.1
+ NMT-all-bt	470,085	32.9	9.2	23.0	8.4
+ RBMT-all-bt	487,862	37.0	14.4	26.4	**11.0**
+ NMT-all-bt + RBMT-all-bt	932,790	38.8	10.9	26.3	9.6
+ NMT-clean-bt	422,596	34.0	9.8	25.0	8.8
+ RBMT-clean-bt	378,567	36.3	**15.5**	25.6	10.9
+ NMT-clean-bt + RBMT-clean-bt	776,006	38.9	11.3	28.2	10.7
+ NMT-clean-bt + RBMT-all-bt	885,301	**40.1**	10.8	**29.9**	9.9

Table 2: Training data sizes (sentence pairs) and results (in BLEU points) for the fin-sme translation models with two different architectures (Transformer and RNN) using original parallel data (Baseline), augmented data sets with unfiltered and filtered backtranslations (all-bt and clean-bt, resp.) evaluated on the UiT test set and the YLE test set.

produced ones to 353,465 sentence pairs. After concatenation with the parallel data and removal of duplicates in this concatenated set, we are left with 422,596 and 378,567 sentence pairs respectively. Furthermore, another training set is created by merging both the NMT and RBMT backtranslations with the parallel data; this set contains 776,006 sentence pairs. The first column of Table 2 shows the training data sizes of the different configurations.

5 Results

The upper part of Table 2 shows the BLEU scores of the translation models trained with the original parallel data set (baseline) and the unfiltered augmented data sets. Similarly to the reverse model, the baseline fin-sme models are well adapted to the UiT test set and do not perform as well with the YLE test set. Adding the NMT backtranslations to the training data gives a significant improvement with respect to BLEU scores: using Transformers on the UiT set, the score raises by 14 points (74% relative), and on the YLE set, the score goes up by 4.9 points (114%). The RBMT backtranslations give an even larger boost on the UiT set than the NMT translations (18.1 points, 96%) and especially on the YLE data (10.1 points, 235%). Using RNNs, the scores are lower overall, but they do show similar improvements with the same training sets as Transformers.

The significant boost from RBMT backtranslations is quite remarkable considering that Apertium does not seem to perform very well on the reverse translation direction on UiT data. This result stresses once more that the effect of backtranslation is to a larger extent due to improved target language coverage than to the quality of the translations. Instead, the additional, less domain-specific knowledge encoded in the RBMT model seems to lead to the additional push even in the UiT domain and it certainly carries over to the out-of-domain data represented by the YLE news data.

The simple combination of both types of backtranslations only provides a modest additional boost on the UiT test set. The out-of-domain performance drops substantially compared to using RBMT-based backtranslations alone. Adding NMT-based translations seem to hurt the model in this regard.

Next, we study the effect of filtering the backtranslations before training the augmented NMT models. Table 2 also shows the results of this approach. We can see that the models benefit from filtering the NMT backtranslations, especially on the UiT domain, whereas the RBMT-based augmentation model performance decreases on the UiT test set. The RBMT-based Transformer model gains an improvement on the YLE set, but the same score with the RNN model decreases slightly. The combination of both backtranslation augmentations leads to a boost in translation quality over the unfiltered backtranslation training set, which suggests that a careful data selection can be important when using data augmentation techniques. The performance on the YLE data is still lower than the RBMT-based data augmentation alone, which could indicate that the RBMT backtranslations are able to carry over out-of-domain information, but this result needs to be taken with a grain of salt as the test set is very small.

Finally, we also train a models that combine filtered NMT backtranslations with unfiltered RBMT backtranslations (last row in Table 2). These models reach the overall highest BLEU scores on the UiT test set, 40.1 with Transformer, but on the YLE test set the performance is lower than with other models, which is a bit surprising but may also depend on random variation and on the small size of the test set.

6 Conclusion

In this work, we confirm that the addition of backtranslations produced with multiple paradigms, including RBMT, improves the quality of NMT models. Additionally, the translation performance can be further improved by removing noisy sentence pairs from the NMT backtranslations. We show that these methods are beneficial in a real-world low-resource setting with the Finnish→Northern Sámi translation pair.

In the future, we plan to extend our work in various ways including more careful data selection and filtering, the use of subword regularization, domain labeling, improved sampling strategies and further data augmentation techniques such as pivot-based translations and transfer learning using multilingual NMT models. Furthermore, we would like to optimize hyper-parameters such as vocabulary size, network architectures and training parameters to maximize the translation performance in low-resource scenarios.

Acknowledgements

The research presented in this paper was supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement № 825627(ELG). This work was also supported by the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement № 771113. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

Lene Antonsen, Ciprian Gerstenberger, Maja Kappfjell, Sandra Nystø Rahka, Marja-Liisa Olthuis, Trond Trosterud, and Francis Tyers. 2017. Machine translation with North Saami as a pivot language. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 123–131.

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.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Meghan Dowling, Teresa Lynn, and Andy Way. 2019. Leveraging backtranslation to improve machine translation for Gaelic languages. In *Proceedings of the Celtic Language Technology Workshop*, pages 58–62.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2021. Transfer learning and subword sampling for asymmetric-resource one-to-many neural translation. *Machine Translation*, pages 1–36.

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.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tommi Pirinen, Francis M. Tyers, Trond Trosterud, Ryan Johnson, Kevin Unhammer, and Tiina Puolakainen. 2017. North-Sámi to Finnish rule-based machine translation system. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 115–122, Gothenburg, Sweden. Association for Computational Linguistics.

Alberto Poncelas, Maja Popović, Dimitar Shterionov, Gideon Maillette De Buy Wenniger, and Andy Way. 2019. Combining SMT and NMT back-translated data for efficient NMT. In *12th International Conference on Recent Advances in Natural Language Processing, RANLP 2019*, pages 922–931. Incoma Ltd., Shoumen, Bulgaria.

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.

Xabier Soto, Olatz Perez-De-Viñaspre, Maite Oronoz, and Gorka Labaka. 2019. Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 8–18, Dublin, Ireland. European Association for Machine Translation.

Xabier Soto, Dimitar Shterionov, Alberto Poncelas, and Andy Way. 2020. Selecting backtranslated data from multiple sources for improved neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3898–3908. Association for Computational Linguistics.

Trond Trosterud and Kevin Brubeck Unhammer. 2012. Evaluating North Sámi to Norwegian assimilation RBMT. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Raúl Vázquez, Umut Sulubacak, and Jörg Tiedemann. 2019. The University of Helsinki submission to the WMT19 parallel corpus filtering task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 294–300, Florence, Italy. Association for Computational Linguistics.