UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Maali Tars

Improving translation for low-resource Finno-Ugric languages with Neural Machine Translation models

Bachelor's Thesis (9 ECTS)

Supervisor: Andre Tättar, MSc

Tartu 2021

Improving translation for low-resource Finno-Ugric languages with Neural Machine Translation models

Abstract:
Training a good neural machine translation model requires a lot of data. The majority of languages in the world have low amounts of suitable data available for this task. One possible solution to this problem is developing a multilingual model, combining high-resource and low-resource languages and creating a shared vocabulary space, where knowledge gained from high-resource languages is applied to translating low-resource languages. Another useful technique is to produce new data for low-resource languages by creating synthetic translations of monolingual data with a baseline model. In this thesis we use both of those methods, training a multilingual baseline model on Finno-Ugric language family data and increasing the amount of data for smaller Finno-Ugric languages by translating monolingual data with the multilingual baseline model in order to improve machine translation quality for low-resource languages.

Keywords: neural networks, automatic learning, machine translation, language technology

CERCS:
P176 Artificial intelligence

Soome-ugri väikeste keelte neuromasintõlke edendamine

Lühikokkuvõte:
Hea neuromasintõlke mudeli treenimiseks on vaja palju andmeid. Väiksema levikuga keeltele aga ei leidu tavaliselt piisavalt andmeid, et treenida mudelit, mis on sama kvaliteetne kui paljude andmetega treenitud mudelid. Üks lahendus sellele probleemile on treenida mitmekeelne mudel, kus on koos paljude andmete ja väheste andmetega keeled, luues sellega ühise sõnavara ning õppimise ruumi. Niimoodi õpivad mudelid tõlkima väiksema levikuga keeli rohkete andmetega keelte abil. Teine võimalus selle probleemi lahendamiseks on kasutada meetodit, kus tõlgitakse ühekeelsed andmed baasmudeli abil teise keelde. Tulemuseks on sünteetilised kahekeelsed andmed, mida saab kasutada uue mudeli treenimiseks. Siin töös treenime soome-ugri keeltega mitmekeelse mudeli ning toodame selle peal ühekeelsetest andmetest sünteetilisi andmeid, et edendada väikese andmemahuga soome-ugri keelte masintõlget.

Võtmesõnad: tehisnärvivõrgud, tehisõpe, raaltõlge, keeletehnoloogia

CERCS:
P176 Tehisintellekt

Contents

1 Introduction

The Finno-Ugric language family has a lot of languages that only a small amount of people still continue to speak. That means there are less and less people who continue to understand it and will potentially carry on speaking them in the future. One way to keep these languages from disappearing is creating translation models, which preserve the structure and vocabulary of the languages and can assist people in learning the languages or make understanding them easier.

In this work we develop a multilingual neural machine translation (NMT) model for five Finno-Ugric language family languages: Estonian, Finnish, Võro, North Saami and South Saami. The aim is to find out whether combining closely related high-resource and low-resource languages together into one model could help improve translation results for the low-resource languages.

Firstly, we develop a baseline multilingual model with parallel data, which consists of sentence pairs from five different language pairings. It has been shown that training a machine translation model, which combines languages that have a lot of parallel data and languages that have low amounts of parallel data, improves the translation quality for the low-resource languages [6]. Also, improving even more so when the languages are closely related to each other [22]. When high-resource and low-resource languages are trained together, they share a vocabulary space and a training space, which enables the model to apply knowledge learned from high-resource languages onto the low-resource language translation process.

As parallel data is scarce for smaller languages, we then use the method called back-translation, where monolingual data is translated to another Finno-Ugric language, which is being used in the model, with the previously trained baseline multilingual model. This results in new synthetic (as opposed to human-translated) parallel data to help train a better NMT model. There is often a lot more monolingual than parallel data available for low-resource languages. In result of the experiments with back-translations, we find out if the technique helps to achieve better results for low-resource language translations.

In addition to previous methods, we perform transfer learning experiments to find out if low-resource language models benefit from being initialized with weights from a model that was trained with large amounts of data from a related high-resource language pair.

In Section 2 there is an overview of the background of this topic and main terminology. Furthermore, we introduce the tools used for developing the model.

Section 3 introduces works that have already been done in this field, which include experiments with low-resource languages using back-translation or multilinguality. Section 4 outlines the main points of the baseline NMT approach and explains differences that were brought into the development with this thesis.

Section 5 sets up the experiments during the development process and gives a more detailed insight into the data that was used for training the models. Finally, in Section

6 there is an analysis of the results on both quantitative and qualitative sides, with quantitative analysis done on the BLEU score, which is an evaluation method for machine translation accuracy [12]. The bigger the score is, the better is the translation accuracy.

The best model that was developed during the experiments (*Para + BT1 + BT2(*)* in Table 4) was deployed on the TartuNLP webpage, where it can be tested by translating text between all the five languages mentioned above [16].

2 Technical background

2.1 Byte pair encoding

Byte pair encoding (BPE) is an algorithm, which is used to divide words into subwords based on learned patterns. It replaces common pairs of byte sequences with a byte that is not used in the data and creates a table of these corresponding symbol bytes and their meanings [2].

Multilingual models especially have big vocabularies and applying this algorithm onto the data helps keep the vocabulary size from growing too big by allowing to determine a fixed-size vocabulary for a model. This results in a vocabulary, where many word segments overlap between languages, which can help train better models, especially for low-resource languages. Additionally, by segmenting words into subwords, this method helps models be creative in translating out-of-vocabulary words, words that the model has never seen before, by combining different word segments [15].

2.2 Transformer architecture

Transformer has become the more widely used architecture type for developing neural machine translation models, which give state-of-the-art results on different translation tasks.

As described by Vaswani et al. [21], it is based on an encoder-decoder network system, but differs from previous architectures by abandoning recurrent (RNN) and convolutional (CNN) layers and is instead exploiting attention mechanisms in their place.

The norm before Transformer architecture was to use mainly RNN layers. But that technique is not equipped to deal with sequences or sentences that are very long. With recurrency, long sequences increase training time significantly, because in order to learn dependencies of a state in the sequence, they need to gather information about the placement and content of the previous states in the sequence. And this results in a lot of extra steps for the learning process, which increases training time.

Transformer works solely on the attention mechanism, which does not spend time acquiring info about the position of a state in the sequence by going through all the previous states. Instead of traversing long paths to remember all the info from previous states, the attention method can decide to focus on the states that are most likely to have the important information needed, concerning the current state. At the start, the input embeddings are given positional encodings, which are stored, to make up for not getting positional info about the states during the learning process.

Attention, more specifically, self-attention, is a function to which you have to give three parameters: queries, keys, values. The output is calculated according to the compatibility of the query with the keys, by summing the values. The produced outputs are then fed to the subsequent encoder/decoder layer. The attention function itself is used

by forming parallel layers in the Transformer architecture. Presuming we have multiple encoder-decoder layers, attention is deployed in three ways: in the encoder-decoder direction, encoder_{i-1}-encoder_i direction and decoder_{i-1}-decoder_i direction.

Removing recurrent and convolutional layers results in some important improvements:

- With attention, NMT models are more parallelizable, have more layers that can learn at the same time, in turn reducing the training time.
- Training time is not dependent on the amount of long sequences in the data, which reduces complexity.
- Removing complex RNN and CNN layers from the architecture makes it easier to implement NMT models.

2.3 Back-translation

Back-translation (BT) is a method where monolingual data is translated using a previously trained baseline model. This creates new translations, which then get paired with the monolingual data, creating a new synthetic parallel data set. For the data set to become back-translation, it is turned around: the synthetic translations become the source side and the monolingual data becomes the target side of the data [14]. This assures that the model learns to translate the source sentence into the target sentence on correct target data, because the synthetic translations could be very flawed and inaccurate, which hurts translation quality. The new data set can then be added to the original parallel data and NMT quality could improve substantially, especially if parallel data for the language or a domain is scarce and monolingual data is in abundance.

Usually, bilingual NMT models are trained in *source* $\rightarrow$ *target* direction. But to use back-translation, there would have to be two models trained: one in *source* $\rightarrow$ *target* and another in *target* $\rightarrow$ *source* direction. One model is needed for producing the data and the other one is trained on the synthetic data [14]. In case of multilingual models however, back-translation is possible with only one model, it produces the synthetic data and also uses it to get better translation results.

2.3.1 Forward-translation

Forward-translation (FT) is also a method for producing synthetic data from monolingual data, but here synthetic translations stay as the target side. Logically, forward-translation is not as efficient for improving translation model results as back-translation. The reason for this is that the target sentences are most likely to be full of mistakes and the model would then learn to translate a source sentence into something that was learned on faulty

data. However, there are experiments in this work that use forward-translations as well as back-translations for testing out this theory.

2.4 Transfer Learning

In the context of NMT, transfer learning is a method where previously trained model parameters are transferred onto another model in a task which should be related to the task that the previous model was trained for [11]. The parameters are initialized for a model before the beginning of the training process with the previous model's weights. If the weights are favourable towards the task at hand, this method enhances the results and overall it saves training time, because the model does not have to start training the parameters from the very beginning .

For low-resource language pairs, this is a favourable technique to improve translation results. Here we firstly train a baseline model on a high-resource parallel data language pair and a multilingual baseline model, where there are high-resource and low-resource language pairs combined. This means a lot of information that is needed for translating the languages involved and the ones closely related to them is learned in the course of training these baseline models. Following that, we can fine-tune these models on the selected low-resource language pair data using the valuable knowledge from previously trained weights and thus save time on training. This technique works best if the languages involved are closely related [22].

3 Related work

3.1 Multilingual models

Multilingual models can be useful for working with low-resource languages in case of NMT. This has been shown in a paper by Gu et al. [6], where the experiments were done combining data from both high-resource and low-resource languages. Low-resource languages can utilize knowledge gained from high-resource languages and as a result, the translation quality of low-resource languages improves significantly. In the paper, there were multiple source languages and one target language. Experiments were done with different sets of languages. The languages that were jointly trained were from various language families, but some of them could be grouped into larger families, for example Romanian, Spanish, French, Italian and Portuguese are all Romance languages. And they note that Romanian achieved better results when the other languages in the model were related to it. In the results, they reported a gain of 5.07 BLEU points on shared lexical representation and a 7.98 BLEU point gain by adding shared representations on a sentence level and back-translations.

The aspect that improves the results is a shared encoding space on a lexical and sentence level. On a lexical level, words are divided into subwords following the BPE algorithm, which helps to identify similar or exactly the same word segments between languages [15]. BPE creates a joint encoding space on chosen languages and the space has more overlapping subwords if the languages are closely related. On a sentence level, languages often share the same syntactic order, for example the principle of whether the verb comes before the subject or after. So low-resource languages can adopt the patterns of a similar higher-resource language.

In a paper by Goyal et al. [5] they added the nuance of choosing languages that were closely related for training a multilingual model and were focused on working with only related languages. The languages originated from Indo-Aryan and Dravidian language families and the objective was to improve translation into English. The data sets used all had an Indian or Hindu language as the source, and English as the target language. In addition, they used transliteration to remove incompatibility among characters between different Indian languages in order to create a common subword space. The average gain overall between the experiments was 5 BLEU points. They reported results in both directions: Indian to English and English to Indian, but also noted that basic multilingual models by themselves did not improve over baseline results. The 5 BLEU point gain came after fine-tuning the multilingual model on a selected language pair. They cite problems and ambiguities in the data to be the reason behind this.

3.2 Back-translation

In the aforementioned paper by Gu et al. [6], back-translation is used as an additional method incorporated into multilingual models to direct the models towards learning more about translating low-resource languages. They utilized a Romanian monolingual corpus, translating it with a pre-trained model, creating a Romanian-English synthetic data set. They applied back-translations to a multilingual model trained on Romanian-English data and observed an improvement of about 3 BLEU points from the baseline score. Even further improvements were made by adding a shared word representation layer to multilingual NMT, which resulted in a gain of about 7 BLEU points.

In another work, by Sennrich et al. [14] they performed back-translation by having two separate bilingual models. One of the experiments was on English to German (EN-DE) and German to English (DE-EN) directions. This way, the models provide back-translated data to each other, which creates a kind of cycle of producing synthetic data followed by additional training. In the additional training phase, the original parallel data was combined with the synthetic translations. German and English are considered to be high-resource languages, which shows in the number of monolingual sentences they had available. Initially, they had 3.6 million sentences in German which they started the back-translation cycles with, translating them on the baseline DE-EN model. They reported improvements made by adding back-translations in both directions by about +2.8–3.4 BLEU points. Additionally, they show that back-translation can be used to improve fluency and to fine-tune for a specific domain.

3.3 Transfer Learning

Transfer learning for low-resource languages has proven beneficial in a work by Zoph et al. [22], where they trained a bilingual parent model on a high-resource language pair and then initialized a bilingual child model, which was going to be trained on a low-resource language pair, using the weights taken from the parent model. The low-resource languages used in those experiments were Hausa, Turkish, Uzbek and Urdu. In addition, they chose which parameters are fixed at the start of training a child model. They used a French-English NMT model as the parent model and separately trained models, initialized with the parent model weights, for the four aforementioned languages: Hausa, Turkish, Uzbek and Urdu. The results showed to have improved the translation quality by a lot for all of the low-resource languages used in the experiments, reporting a BLEU score gain of 5.6 points compared to baselines.

With additional experiments, they also found that it matters how similar the languages used in the parent and child models are. While developing a child model for low-resource Spanish-English task, using a French-English parent improved the results more than using a German-English parent, the difference being about 1 BLEU point. Spanish is very similar to French as they both originate from the Romance language family, whereas

German is from the Germanic language family, which is quite different, but as both of these language family languages are spoken in geographically close areas, there are still some similarities.

4 Approach

4.1 Baseline approach

Nowadays, neural machine translation is the preferred approach for solving translation tasks. The best results for low-resource language models have been achieved by using statistical machine translation models [22]. But NMT is easier to implement because it generalizes well over different tasks.

Overall, the state-of-the-art results in machine translation tasks have come from exploiting the previously explained Transformer architecture. But the models are usually bilingual and one-directional, meaning there is one source language and one target language. Training is usually done with a lot of parallel data from high-resource languages, for example English, German and Spanish, where amounts of data are in millions of sentence pairs. Low-resource languages do not usually get very good results with NMT models, because of the lack of parallel data in large amounts, which is needed for NMT models to give good results. Experiments in the paper by Gu et al. [6] (Figure 1), show that NMT starts producing reasonable results, when there is at least 13 000 sentence pairs, but gives poor results when the number is under that or there is no data available. The experiment shows that up to 13 000 sentence pairs, BLEU score stays below 5. But training with 28 000 sentence pairs makes the jump to about 12 BLEU points.

When dealing with languages that are popular in the world, high-resource could mean millions of sentence pairs. But in the context of this thesis, where we deal with the smaller and less popular Finno-Ugric language family, a high-resource language is defined as a language that has over 100 000 sentence pairs and a low-resource language has less than 100 000 sentence pairs.

4.2 Enhanced approach

In this work we attempt to improve the results of translating low-resource languages. For that we employ the help of related languages that come from the same language family, both high-resource and low-resource languages. Some languages are more closely related than others and they share a number of syntactic and lexical similarities.

Firstly, we learn text patterns from a related language that has a lot of parallel data, which is the high-resource language. This information can then be used for learning to translate the low-resource languages, because as the languages are related, some of the patterns apply to both of the languages.

In the experiments of this work, we use the Finno-Ugric language family. When creating a multilingual NMT model, we exploit two more frequently spoken languages, Finnish and Estonian, and build a shared embedding space adding Võro, North Saami and South Saami languages to Finnish and Estonian. Since Võro is spoken in Estonia, mainly in South-Estonia, Estonian and Võro are very closely related languages, which means

that sentence syntactics and much of the words are the same or can be logically derived from one another. So patterns learned from Estonian can be used to translate or learn how to translate Võro language. Also, Finnish and both Saami languages are geographically very close to each other, with Finnish and North Saami being grammatically more similar to each other than Finnish and South Saami, which are quite different.

Additionally, we make use of monolingual data, which in case of low-resource languages can be found in larger numbers than parallel data. We use the back-translation technique, which helps produce synthetic translations of monolingual data using the multilingual baseline model. The paired up synthetic and monolingual data is then used as added parallel data for improving on baseline results. This requires training another model, but this time including the synthesized parallel corpora. As the focus is on improving the results of low-resource language translations, we down-sample the high-resource language monolingual data, in this case Finnish and Estonian, by reducing the number of monolingual data to even the numbers between all of the languages. This way the model can concentrate more on the smaller languages.

5 Experiments

5.1 Data

The data used in this work consisted of five Finno-Ugric languages: Estonian (et), Finnish (fi), Võro (vro), North Saami (sme), South Saami (sma). All parallel data and monolingual data is representing different domains and domain-specific experiments were not performed.

5.1.1 Parallel data

There were five different sets of language pairs that had parallel data available for developing this model. They were Estonian-Finnish, Estonian-Võro, Finnish-North Saami, Finnish-South Saami and North Saami-South Saami.

The parallel data sets that were recovered, were in different file formats and contained various levels of incorrect or coarse data, which means the data had to be cleaned and adjusted so that the different sets could become compatible with each other. The cleaned data sets consisted of aligned sentences, where sentences were separated by a line-break. The cleaning process of parallel data included removing empty lines and occasionally having to correct sentence alignments.

The files that contained the data had to firstly be put into a conformed format, which ultimately was chosen to be a simple TXT file format. The parallel data files were mainly of XML, TMX and TXT type and each required a different approach of extracting the sentences from the files and joining them into a TXT file type, because they lacked a conformed syntax, even amongst the XML files. Some of the files had made a distinction between each sentence, but others had a paragraph as the smallest unit. For extracting sentences from paragraphs and setting each sentence on its own line, we utilized a tokenizer from the Natural Language Toolkit (NLTK)[1] that splits text into sentences.

While performing preliminary experiments with all the parallel data, it became clear that some language pairs had more repeated sentences in them than others, which caused the training, testing and developing sets to overlap. This gave a false sense of a good translation quality, because some test sentences were also present in the training or developing data. So there was a need for additional pre-processing of the data sets by performing a check that each sentence was unique and eliminating the repeated sentences within the data set before dividing it into train-test-dev sets. Table 1 gives an overview of how many sentences were removed in the pre-processing step.

In the experiments, the Estonian-Finnish language pair was the high-resource parallel data language pair and other pairs were in significantly lower numbers as can be seen in Table 1. The cleaned Estonian-Finnish data consisted of about 2.6 million sentence

[1] https://www.nltk.org/

Language pair	Before cleaning	After cleaning	Eliminated sentences
et-fi [17]	3 566 826	2 646 922	919 904
et-vro [7]	30 816	30 502	314
fi-sme [20]	109 852	35 426	74 426
fi-sma [20]	3098	2895	203
sme-sma [20]	23 746	21 557	2189
Overall	3 734 338	2 737 302	997 036

Table 1. Parallel data sets (in sentence pairs)

Language pair	train	test	dev
et-fi	2 646 371	362	189
et-vro	29 902	400	200
fi-sme	34 826	400	200
fi-sma	2445	300	150
sme-sma	20 957	400	200
Overall	2 734 501	1862	939

Table 2. Parallel data train-test-dev distribution (in sentence pairs)

pairs, but other pairs made up only about 1 percent of the entire parallel data set with Finnish-South Saami occupying a very low number of about 0,1 percent of the data set.

After cleaning the data, it was divided into training, testing and developing data sets. The complete overview of the size of each set per language pair can be found in Table 2. Test and dev set sizes were quite small because there is already not a lot of parallel data for the low-resource language pairs available, so this way the majority stays in the training set. For the experiments in this work, most of the language pairs have about 400 sentences in the testing set and about 200 sentences in the developing set.

5.1.2 Monolingual data

Monolingual data was available for all five languages. Naturally, Estonian and Finnish had data in abundance, but for the sake of the models learning more about low-resource languages, we used the down-sampling technique, which means reducing the amount of Estonian and Finnish monolingual data to bring them on the same level with the amount of low-resource language monolingual data in use. This way the model is not mainly focusing on learning Estonian and Finnish.

There were two rounds of back-translation, meaning that the entire monolingual data set was divided into two sets. Both sets are described in Table 3. The first set consisted of 100 000 sentences of Estonian and Finnish with Võro having the most data and North

Language	First set	Second set	All
et [13]	100 000	25 000	125 000
fi [4],[9]	100 000	25 000	125 000
vro [8], [19]	162 807	5290	168 097
sme [18],[4],[1],[10]	33 964	6057	40 021
sma [18],[3],[1]	55 088	5377	60 465

Table 3. Monolingual data sets.

Saami and South Saami having about 34 000 and 55 000 sentences, respectively. The second set had lower numbers of data, Estonian and Finnish both with 25 000 sentences while the other languages had data in the range of about 5000-6000 sentences.

Pre-processing the data was a long process as there were no conclusive ready-made sets available for languages like Võro, North Saami and South Saami. Among the first set, the data was composed mostly from news corpuses, fiction and Wikipedia texts and the data files were in different formats, as was the case with parallel data. Estonian and Võro required extracting sentences from texts and removing empty lines. Therefore, most of the methods used for parallel data were employed on the monolingual data as well.

The Võro, North Saami and South Saami data in the second set was gathered manually from news articles and various PDF style documents available. The data then needed to be broken into sentences and joined into one TXT type file for compatibility.

5.2 General experiment settings

All the models in the experiments were trained using the Sockeye[2] framework, which implements the state-of-the-art Transformer architecture. Prior to training, the data was binarized and optimized with the ready-made method from Sockeye, otherwise this step would be done repeatedly during training, which increases training time. In terms of vocabulary settings during training, a shared vocabulary setting was used, which creates a joint vocabulary between source and target data. Batch size was set to 6000, checkpoint interval to 2000. The maximum number of checkpoints on which the quality of the model does not improve was set to 32. Training ends when the model reaches those 32 consecutive checkpoints. The models were given 7 days for training, but all the models trained until the 32 checkpoints were passed, with some training a bit longer than 7 days. For multilingual models, it is required that the model is given some info about which language the sentences should be translated into. For that the model also takes in a source factor file, which contains that information in token form.

[2]https://awslabs.github.io/sockeye/index.html

For applying the BPE algorithm onto the data we used Sentencepiece[3], which is a text tokenizer and detokenizer. A Sentencepiece model is usually trained on the training data that the NMT model is going to be trained on or on data that is similar to it, so the tokenizer can learn the patterns that are found in the training data. After that, the Sentencepiece model takes untokenized sentences from the training data and with the help of previously learned patterns, it encodes each sentence, cutting the words into subwords where it deems suitable.

5.3 Baseline models

Baseline models for each language pair are necessary for comparing them to the results of the enhanced approach experiments, to work out whether the translation capabilities improved for the low-resource languages or not.

In this work, the baseline models that were developed were bilingual and one-directional, meaning ten baseline models were trained, each in a different direction. Because there were five language pairs making up the parallel data, the models were as follows in the *source* $\rightarrow$ *target* direction: et-fi, fi-et, et-vro, vro-et, fi-sme, sme-fi, fi-sma, sma-fi, sme-sma, sma-sme.

Each baseline model was trained with the parallel data of that language pair going only in one direction, one source language and one target language. For each of the baseline models, the data was beforehand tokenized by a separate Sentencepiece tokenization model. The tokenization model was trained on the same parallel data that was later used to train the baseline model. This means that each baseline model had its own shared embedding space, limited to the two languages used in training the baseline model.

5.4 Multilingual baseline model

One of the main experiments of this work was developing the multilingual baseline model, which had five source languages and five target languages. This means that one model could produce translations in 20 different directions. This is the experiment which, compared to the bilingual baseline models, will show the benefits of combining high-resource and low-resource languages into one model, highlighting how the model can use knowledge from high-resource languages and apply that when translating low-resource languages.

Firstly, the multilingual model was trained on all of the parallel data. To make it have 20 directions, each pair of parallel data was copied and source-target direction was switched. The turned-around parallel data set was then added to the original data set. For example, Estonian-Finnish was copied and reversed, which created the Finnish-Estonian data set. All ten parallel data directions were then combined into one data set, doubling

[3]https://github.com/google/sentencepiece

the number of data seen in Table 1, going from about 2,7 million sentence pairs to 5,4 million sentence pairs. All of the data was used to train that one multilingual model.

The data was tokenized with a Sentencepiece model, similarly to the baselines. For the multilingual baseline model, however, the tokenization model had all five languages in the same embedding space instead of two languages, like the bilingual baseline models had. This means the tokenization patterns, that all the parallel data was tokenized by, were different from the patterns used to tokenize bilingual baseline models. In the multilingual tokenization model, the patterns were more likely to be generalized over all five languages rather than being specific to two languages.

5.5 Back-translation experiments

Back-translation models, in comparison with multilingual baseline, will show the gains that come from utilizing monolingual data for enhancing translation results for low-resource languages. There were two cycles performed on back-translation data. During both cycles, the monolingual data was translated into every other language in equal measures. For example, referring to Table 3, 1/4 of the 100 000 sentences in Estonian were translated into Finnish, another 1/4 into Võro, another 1/4 into North Saami and the last 1/4 of sentences into the South Saami language.

The result is synthetic translations of monolingual data, that when paired together with the monolingual data, make up an additional parallel data corpus. Knowing that synthetic translations might be flawed and inaccurate, the data was turned around - the synthetic data became the source language and the monolingual data became the target language, making it back-translation. This technique can then help the model give more accurate translations, because it is learning on correct sentences rather than on the generated synthetic data that might not be entirely accurate.

The new synthetic parallel data corpus is added onto the original, human-translated corpus, which gives the model more data to learn and generalize off of. For better overview and better results, we performed two cycles of back-translation. Mostly, the techniques of both cycles were the same, but there were some differences:

1. The first batch of synthetic data was produced with the multilingual baseline model that was trained only on human-translated parallel data, which was described in the previous section. The synthetic data was then added to the original parallel data and the training process was repeated, which produced a new model.

2. Following that, the process of creating synthetic data was repeated, only now the monolingual data was translated by the newest model that had been trained on parallel data and synthetic data from the first cycle of back-translation. Subsequently, a new model was trained using parallel data plus the two batches of synthetic data produced.

5.5.1 First back-translation cycle

The first model with back-translation data was developed using multilingual baseline model data and it had more monolingual data than the second monolingual data set. The weights of parameters that were trained during the multilingual model training process, were transferred onto the new model before training, which saves training time. As choosing the ratios of monolingual data can guide the learning process towards a specific language, in the first cycle, the training should have been turned towards learning to translate into Võro, because there was significantly more data on Võro language than other languages, as can be seen in Table 3. In addition to the main model with parallel data and back-translation data on multilingual model weights, there were a number of side-experiments performed to determine the beneficial aspects of the enhanced approach:

- A model with all original parallel data and data from the first back-translation cycle without using multilingual baseline model weights.
- A model with only back-translation data 1) on multilingual baseline model weights, 2) without multilingual baseline model weights.
- Experiment combining back-translation and forward-translations into the parallel data corpus 1) on multilingual baseline model weights, 2) without multilingual baseline model weights. For testing if back-translation is a more effective technique to use.

5.5.2 Second back-translation cycle

The second model with back-translation was developed starting with the weights from the last cycle's model. The second monolingual data set had a lot less data than the first one. The data coming from the second back-translation cycle consisted of 1) shuffled and back-translated (now by the model that was trained in the last cycle) first monolingual data set and 2) back-translated second monolingual data set. The data from the first and second back-translation cycles was combined and used to train the new model. Additional experiments were as follows:

- A model with all original parallel data and all synthetic data from both back-translation cycles without using any pre-trained weights.
- Experiment combining first and second cycle's back-translations and forward-translations with the parallel data corpus 1) on last cycle's model weights, 2) without using any pre-trained weights.
- A model with parallel data and only data from the second back-translation cycle.

5.6 Transfer Learning experiments

Another way to enhance results for low-resource languages in NMT, is the transfer learning approach. The aspects of this technique were already used in developing the models with back-translation data that used the parameter weights from the previously trained models. However, that was for improving the results for all five languages at once.

In this work we performed an experiment fine-tuning the multilingual baseline model and an et-fi baseline model on et-vro parallel data. Then we compared the results to the et-vro baseline model results. And then compared the results between the fine-tuning on either of the baseline models, to see where the benefits come from.

The et-vro data for fine-tuning was the same parallel data that was used for training the multilingual model. This way the model is more focused on learning to translate Estonian to Võro.

6 Results

For testing the models, there was a test set of parallel data for each language pair, which is described in the data section. The source side of the parallel data was translated using the model that was being tested. The translations were then compared to the original target side sentences. Translation took place using the Sockeye framework.

6.1 Quantitative analysis

For quick quantitative comparison between translation proficiency we used the BLEU score, which is an evaluation method for machine translation accuracy [12]. The hypothetic translations and the expected translations were first detokenized, changed back into human-readable state. Then the hypothetic-expected translation pairs were given as parameters to a BLEU method calculation implementation SacreBLEU[4]. It helps get BLEU results quickly and easily. BLEU score is a positive number and the bigger the number, the better the score.

The same test set was used for all models, so the scores do not depend on differences in the testing data. The scores for all the best experiments are detailed in Table 4. The table describes the best models and also weight transfer analysis comparing two models trained only on back-translations. There is significant growth for all the low-resource language pairs, when comparing baseline scores to the multilingual baseline (ML) model scores. The average gain was 7.6 BLEU points, with the highest gains coming from Võro to Estonian (Δ=12.1) and North Saami to South Saami (Δ=11.5). Lowest gain, but still a significant one, was for Finnish and South Saami pair, perhaps because there was a lot less data for that low-language pair with only about 3000 sentence pairs, as can be seen in Table 1. The high-resource language pair scores got worse by about 0.5 BLEU points. However, that is a marginal drop and an acceptable trade-off, because the aim of the work was to enhance translation results for low-resource languages, which was a goal that the multilingual model achieved.

The biggest improvements from the experiments with back-translation data are noticeable from the model with parallel data and back-translations but without pre-trained parameters. Seems that in this case the pre-trained weights mostly hindered the learning process for low-resource language pairs. However, the differences in scores between training with pre-trained weights and with no weights are still quite marginal.

Adding forward-translation sets to parallel data and back-translation data sets did not improve the scores except for the Estonian and Võro pair.

Table 4 also describes experiments that included back-translations of 1) the second monolingual data set and 2) the shuffled and re-translated first monolingual data set, which both were produced by the *Para + BT1* model. *BT1* signifies the back-translations

[4]https://github.com/mjpost/sacrebleu

of the first set of monolingual data. *BT2* signifies the first and second set of monolingual data sets, but the first set was translated differently from the last back-translation cycle in that the sentences were shuffled and mostly translated into another language than they were in *BT1*.

With the data from the second back-translation cycle added to models, the results vary between different combinations of data and pre-trained parameters, so none could seemingly be chosen as the best model. The biggest improvements can be seen from the *Para + BT1 + BT2(*)* model, that increased et-vro, fi-sma, sme-sma BLEU scores. This model was also chosen to be deployed on the TartuNLP webpage for testing and translating text in all the 20 different directions [16].

As we see from parallel data and back-translation experiments, the pre-trained weights taken from the multilingual baseline model did not help all of the models which were trained on both data sets. But the experiment done with only back-translation data still proves that transferring pre-trained weights can help achieve better results for a translation model. The difference between the two models trained on solely back-translation data is on average 7.1 BLEU points.

Additional transfer learning experiments showed that fine-tuning a previously trained model on a low-resource language pair increases the scores even more than back-translation combined with multilinguality could achieve. There were two transfer learning experiments, one with et-fi model fine-tuned on et-vro and one, where the multilingual baseline was fine-tuned on et-vro. Comparing the results to each other in Table 5, it is clear that the benefits come mostly from having et-fi language pair in the multilingual baseline model, because adding other language pairs into the mix increased the score only by 1 BLEU point, whereas having only et-fi model parameters can give a 12 BLEU point boost. But transfer learning alone has a downside, because in the context of this work, it means having to train five or ten separate models. The multilingual model method is more compact, because you can train one model and have 20 different directions to translate into. With transfer learning there would be a number of different models, which is not as comfortable nor efficient.

Model	et-fi	fi-et	et-vro	vro-et	fi-sme	sme-fi	fi-sma	sma-fi	sme-sma	sma-sme	Avg
Baselines	32.0	29.4	14.6	17.5	28.0	28.7	4.6	6.3	8.3	9.1	
Multilingual (ML)	30.9	29.5	23.8	29.6	31.3	34.7	9.4	9.4	19.8	19.8	7.6
Para + BT1	**32.4**	29.9	25.2	29.4	**32.3**	36.1	10.8	9.9	20.3	20.0	8.4
Para + BT1(*)	30.1	29.1	24.5	30.3	32.3	**36.2**	**11.1**	**10.5**	**21.4**	20.0	8.7
Para + BT1 + FT1	31.3	**30.1**	25.2	**31.5**	31.3	35.7	8.9	10.0	18.7	**20.4**	8.1
Para + BT1 + FT1(*)	30.9	28.8	**25.8**	30.4	31.5	35.7	8.9	10.1	19.4	20.1	8.1
Para + BT2	31.5	**30.2**	26.0	31.0	32.3	36.6	11.3	**10.9**	20.3	**21.0**	9.0
Para + BT1 + BT2(*)	31.3	29.6	**26.2**	31.3	31.4	36.4	**12.4**	10.6	**21.6**	20.7	9.2
Para + BT1 + BT2(**)	30.4	29.7	25.1	31.6	31.7	**37.5**	11.4	10.3	21.3	20.9	9.1
Para + BT1&2 + FT1&2(*)	30.2	29.4	25.1	**31.7**	31.5	36.8	9.5	9.7	20.4	20.6	8.5
BT1	21.1	21.6	20.5	24.9	24.0	27.4	8.5	7.3	15.9	14.5	3.3
BT1(*)	8.4	8.6	18.7	19.9	11.8	13.4	6.9	5.3	12.9	9.2	-2.4

Table 4. BLEU scores. (*) - trained without pre-trained weights, (**) - trained on *Para + BT1(*)* weights. *Para* - original parallel data set, *BT* - back-translation data set, *FT* - forward-translation data set, *Avg* - average gain over baselines (excluding et-fi and fi-et).

Model	et-vro
baseline	14.6
et-fi fine-tuned on et-vro	26.5
multilingual baseline fine-tuned on et-vro	27.6

Table 5. BLEU scores for transfer learning experiments

6.2 Qualitative analysis

Tables in the appendix show the qualitative analysis performed for example translations of Võro to Estonian and Estonian to Võro directions. Comparing baseline results to other models, the results generally make a big improvement from baseline to multilingual. And models with back-translation improve even more, making some important changes for understanding the context.

6.2.1 Võro to Estonian

Table 6 in the appendix shows translations from Võro to Estonian between different models. Some analysis with concrete examples:

- **Source:** Nuuri om umajago pääle kasunuq, **vaest** võtt kiäki nuur küläelo vidämise üle.
- **Reference:** Noori on omajagu peale kasvanud, **vast** võtab keegi noor külaelu vedamise üle.
- **Best translation:** Noori on omajagu peale kasvanud , ehk võtab keegi noor külaelu vedamise üle .

The first sentence in Table 6 is practically translated correctly, only the word "vast" has been replaced by "ehk", but they have the same meaning and both work in this context.

- **Source:** Parhilla ommaq jutuq hindamis**kogo** käen, kokko**võtõq** ja preemiäsaajaq trüki**täseq** ärq järgmädsen **Uman** Lehen.
- **Reference:** Praegu on jutud hindamis**komisjoni** käes, kokku**võte** ja preemiasaajad trüki**takse** ära järgmises **Uma** Lehes.
- **Best translation:** Praegu on jutud hindamiskogu käes , kokkuvõtted ja preemiasaajad trükivad ära järgmises Uman Lehes .

This translation has some bigger flaws. It shows that the model is not very good at handling names, with "Uman Lehen" being translated incorrectly to "Uman Lehes". "kokkovõtõq" is translated to plurality, but it should be singular. In addition, "trükitäseq" is translated to the wrong verb form: "-vad" should be replaced with "-takse". "hindamiskomisjon" and "hindamiskogu" basically mean the same thing, but "hindamiskomisjon" is a more common term.

- **Source:** Võrokõsõ välläkutsõ **mäng härgüt' kõnõlõmist** pruuvma.
- **Reference:** Võrokese väljakutse **mäng ärgitas kõnelemist** proovima.
- **Best translation:** Võrukese väljakutsele mängib innukalt kõnelusi proovima.

This translation is quite bad and it is unclear what is meant by the sentence. The model confuses a noun with a verb for the words "mäng" and "mängib" and also for the words "kõnõlõmist" and "kõnelusi". The model has replaced nouns with respective verbs and that makes the sentence non-sensical.

Additional examples and translations from other models are displayed in Table 6.

6.2.2 Estonian to Võro

Table 7 in the appendix gives examples of Estonian to Võro translation.

- **Source:** Ja ühe hea külje leiab siidritegija uue nime juures veel.
- **Reference:** Ja üte hää küle löüd siidritegijä vahtsõ nime man viil.
- **Best translation:** Ja üte hää küle löüd siidritegijä vahtsõ nime man viil.

First sentence is translated correctly.

- **Source:** Uue nime **väljamõtlemisel** oli tähtis, et oleks selge side kohaliku kogukonnaga ja et nimi **aitaks jutustada** ettevõtte lugu.
- **Reference:** Vahtsõ nime **vällämärkmise man** oll' tähtsä, et olõs selge köüdüs paikligu kogokunnaga ja et nimi **avitanuq jutustaq** ettevõtmisõ luku.
- **Best translation:** Vahtsõ nime vällämõtõldõn oll' tähtsä, et olõs selge side paigapäälidse kogokunnaga ja et nimi avitas kõnõlda ettevõttõ lugu.

For the third example in Table 7, the best model decides to use a direct translation of "väljamõtlemisel" to "vällämõtõldõn". In addition, on translating, the model omits the "q" endings of the words "avitanuq" and "jutustaq" or "kõnõldaq". Overall, in a lot of other sentences the "q" endings were not being added by the models, the symbol usually signifying plurality. Although, that might be an issue that stems from the training data, where sometimes words are written without those endings and that may confuse the model.

- **Source:** Leevakul elab **ametlikult** pea 300 inimest.
- **Reference:** Leevakul eläs **kirjo perrä** pia 300 inemist.
- **Best translation:** Leevakul eläs virallisesti pia 300 inemist.

In the sixth sentence, translating the word "ametlikult" ("officially" in English), the best model chooses a Finnish word "virallisesti" instead of a Võro phrase "kirjo perrä".

Additional examples and translations from other models are displayed in Table 7.

6.2.3 Finnish to North Saami

Table 8 in the appendix shows examples of translations from Finnish to North Saami.

- **Source:** Uutta nimeä **keksiessä** oli tärkeää, että olisi selkeä yhteys paikalliseen yhteisöön ja että nimi auttaisi kertomaan yrityksen tarinan.
- **Reference:** Ođđa nama **hutkkadettiin** lea dehálaš, ahte livčče čielga oktavuohta báikkálaš servvodahkii ja ahte namma veahkehivčče muitalit fitnodaga muitalusa.
- **Best translation:** Ođđa nama huksemis lei dehálaš, ahte livčči čielga oktavuohta báikkálaš servodahkii ja ahte namma veahkehivčči muitalit fitnodaga máidnasa.

In the first example, the best model gives an almost perfect translation, but chose the wrong word "huksemis", which approximately means "to build" in English. But the meaning necessary here is for the phrase "coming up with something".

- **Source:** Nuoria on tullut tilalle **aika lailla**, ehkä ottaa eräs nuori kyläelämän vetämisen haltuunsa.
- **Reference:** Nuorat lea boahtán lasi **oalleláhkái**, gánske muhtun nuorra váldá gilieallima geassima iežas háldui.
- **Best translation:** Nuorat leat boahtán sadjái áige ládje, soaitá váldit ovtta nuorra gilieallima jođiheami háldui.

The fourth translation in Table 8 has kept its meaning but the word-pair "áige ládje" is a direct translation from Finnish and does not hold the same meaning as the correct translation does.

- **Source:** Sosiaalisessa mediassa pitivät ihmiset eniten tehtävästä “Puhu tai postaa **yksi vitsi** tai tarina **võron kielellä**”.
- **Reference:** Sosiála medias liikojedje olbmot maiddái bargobihtás “Muital dahje postte **ovtta cukcasa** dahje máidnasa **võro gillii**.”
- **Best translation:** Sosiála medias dolle olbmot eanemusat bargguin “Puhu dahje poasta okta njaš dahje máidnasa vuonagillii”.

In the sixth example, the models had problems with translating the Võro language name. Additionally, "okta" has the wrong case. It should be "ovtta" and the word itself means "one" in English. A little suprisingly, the model could not translate correctly the Finnish word "vitsi", which means "a joke" in English. The best model chooses the word "njaš" instead of the correct word "cukcasa".

An additional problem, that was noticed in other examples, was the use of impersonal style. In North Saami language, impersonal style is not used very often as it makes the sentences very hard to understand. Some of the models, however, occasionally translated words into impersonal style, so the models did not learn that aspect of the language.

7 Conclusion

In this work we developed multiple neural machine translation models in attempt to improve the translation results for small Finno-Ugric language family languages. For that we trained a multilingual model combining both high-resource and low-resource Finno-Ugric languages into the same model. Additionally, we utilized the back-translation method by producing more parallel data for low-resource languages from monolingual data. The experiments that were performed combined these two methods of multilinguality and back-translation. In addition, we did two transfer learning experiments: 1) fine-tuning a bilingual baseline model and 2) fine-tuning the multilingual baseline model.

Comparing the results of baseline models to the enhanced models, it showed that there is a huge benefit to combining high-resource and low-resource languages into one multilingual model. We saw that the models take on the knowledge from high-resource languages and use that to translate low-resource languages. Something that training a bilingual low-resource model does not have.

Adding synthetic parallel data via back-translation is also a useful method, but proved to be more marginally beneficial than a multilingual shared vocabulary space. Transferring parameters from the multilingual baseline model to the models with back-translations, was not an overtly useful detail in training. However, comparing the results of the experiments with only synthetic data and with or without pre-trained parameters, did show a significant difference in the BLEU score, meaning pre-trained parameters can still have a positive effect.

When considering the transfer learning experiments, it became clear that the big improvements came from using a high-resource language pair model as the baseline, in this case the Estonian-Finnish model. Fine-tuning a low-language pair on the multilingual baseline model, however, had an even better effect, although the difference was not very big. In the case of transfer learning and this work however, there would have had to been ten separate models trained, whereas the one multilingual model can translate in all of the 20 possible directions.

For future experiments it would be interesting to add other Finno-Ugric languages into the mix, because there are a lot of different Saami languages. Additional experiments with high-resource languages that are outside the Finno-Ugric language family might also prove to be useful. For example, using English as a high-resource language in training the models or rather German, because of Estonia's history with Baltic German culture.

References

[1] *freecorpus - Revision 10181: /orig*. URL: https://victorio.uit.no/freecorpus/orig/ (visited on 22/11/2020).

[2] Philip Gage. "A New Algorithm for Data Compression". In: *C Users J.* 12.2 (Feb. 1994), pp. 23–38. ISSN: 0898-9788.

[3] UiT The Arctic University of Norway Giellatekno - Saami Language Technology and The Divvun group at UiT The Arctic University of Norway. *SIKOR South Saami free corpus*. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. 2015. URL: http://hdl.handle.net/11509/102 (visited on 14/01/2021).

[4] D. Goldhahn, T. Eckart and U. Quasthoff. "Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages". In: *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*. 2012.

[5] Vikrant Goyal, Sourav Kumar and Dipti Misra Sharma. "Efficient Neural Machine Translation for Low-Resource Languages via Exploiting Related Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Online: Association for Computational Linguistics, July 2020, pp. 162–168. DOI: 10.18653/v1/2020.acl-srw.22. URL: https://www.aclweb.org/anthology/2020.acl-srw.22.

[6] Jiatao Gu et al. *Universal Neural Machine Translation for Extremely Low Resource Languages*. 2018. arXiv: 1802.05368 [cs.CL].

[7] S. Iva. *Võru - eesti paralleelkorpus*. 2019. URL: https://doi.org/10.15155/1-00-0000-0000-0000-001A0L (visited on 14/01/2021).

[8] S. Iva. *Võru ja Setu ilukirjanduskorpus*. 2019. URL: https://doi.org/10.15155/1-00-0000-0000-0000-00186L (visited on 14/01/2021).

[9] Universität Leipzig. *Wortschatz*. 2021. URL: https://wortschatz.uni-leipzig.de/en/download/finnish (visited on 14/01/2021).

[10] Universität Leipzig. *Wortschatz*. 2021. URL: https://wortschatz.uni-leipzig.de/en/download/northern-sami (visited on 14/01/2021).

[11] Emilio Soria Olivas et al. *Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques - 2 Volumes*. Hershey, PA: Information Science Reference - Imprint of: IGI Publishing, 2009. ISBN: 1605667668.

[12] Kishore Papineni et al. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://www.aclweb.org/anthology/P02-1040.

[13] University of Tartu Research Group of Computational Linguistics. *Mixed Corpus of Estonian: Eesti Päevaleht*. 2018. URL: https://www.cl.ut.ee/korpused/segakorpus/epl/ (visited on 25/10/2020).

[14] Rico Sennrich, Barry Haddow and Alexandra Birch. *Improving Neural Machine Translation Models with Monolingual Data*. 2016. arXiv: 1511.06709 [cs.CL].

[15] Rico Sennrich, Barry Haddow and Alexandra Birch. *Neural Machine Translation of Rare Words with Subword Units*. 2016. arXiv: 1508.07909 [cs.CL].

[16] TartuNLP and University of Tartu. *TARTUNLP SOOME-UGRI NEUROTÕLGE*. 2021. URL: https://soome-ugri.neurotolge.ee/ (visited on 14/01/2021).

[17] Jörg Tiedemann. “OPUS – parallel corpora for everyone”. In: *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*. Riga, Latvia: Baltic Journal of Modern Computing, May 2016. URL: https://www.aclweb.org/anthology/2016.eamt-2.8 (visited on 14/01/2021).

[18] Jörg Tiedemann. “Parallel Data, Tools and Interfaces in OPUS”. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012. ISBN: 978-2-9517408-7-7.

[19] *Uma Leht*. 2020. URL: https://umaleht.ee/ (visited on 22/11/2020).

[20] UiT Norgga árktalaš universitehta. *Translation Memory*. Online, 2020. URL: https://giellalt.uit.no/tm/TranslationMemory.html (visited on 13/07/2020).

[21] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[22] Barret Zoph et al. *Transfer Learning for Low-Resource Neural Machine Translation*. 2016. arXiv: 1604.02201 [cs.CL].

Appendix

Source	Nuuri om umajago pääle kasunuq, vaest võtt kiäki nuur küläelo vidämise üle.
baseline	Naaž on omajagu kasvanud, ehk keegi võtab noor külavahelaulu üle.
multilingual	Nüüd on omajagu kasvanud, võib-olla võtab keegi noor külaelu vedamisest üle.
Para+BT1+FT1	Noori on omajagu kasvanud, ehk võtab keegi noor külaelu vedamise üle.
Para+BT1&2+FT1&2(*)	Noori on omajagu peale kasvanud , ehk võtab keegi noor külaelu vedamise üle .
Reference	Noori on omajagu peale kasvanud, vast võtab keegi noor külaelu vedamise üle.

Source	A edesi ei näeq tükk aigo tii pääl üttegi võrokiilset silti.
baseline	Aga edasi ei näinud tükk aega, tee ühtegi saaklooma ära.
multilingual	Aga edasi ei näe tükk aega tee peal ühtegi võrukeelset ikka.
Para+BT1+FT1	Aga edasi ei näe tükk aega tee peal ühtegi võrukeelset silti.
Para+BT1&2+FT1&2(*)	Aga edasi ei näe tükk aega teel ühtegi võrukeelset silti.
Reference	Aga edasi ei näe tee peal tükk aega ühtegi võrukeelset silti.

Source	Ütelt puult tulõ hoita vannu mõtsu, et nä saanu ummi pessi kohegi ehitä.
baseline	Ühis poolt tuleb hoida vanade metsade, et nad saanud märkimisväärset kuhugi ehitanud.
multilingual	Ühel pool tuleb hoida vana metsa, et nad saaksid oma pesu ehitada.
Para+BT1+FT1	Ühel pool tuleb hoida vana metsa, et nad saaksid oma pessi kuhugi ehitada.
Para+BT1&2+FT1&2(*)	ühelt poolt tuleb hoida vanu metsasid, et nad saaksid oma pessi kuhugi ehitada.
Reference	Ühelt poolt tuleb hoida vanu metsi, et nad saaks oma pesasid kuhugi ehitada.

Source	Parhilla ommaq jutuq hindamiskogo käen, kokkovõtõq ja preemiäsaajaq trükitäseq ärq järgmädsen Uman Lehen.
baseline	Praegu on instruktsioone, ka kokkuvõtted, selliste sündmuste ja aegajalt „sisse lülitada" kaugemate Leivalentsemad.
multilingual	Praegu on jutud hindamiskogu käes, kokkuvõtted ja preemiasaajad trükivad järgmise Uman Leheni.
Para+BT1+FT1	Praegu on jutud hindamiskogu käes, kokkuvõtted ja preemiaad trükitavad järgmises Uman Lehes.
Para+BT1&2+FT1&2(*)	Praegu on jutud hindamiskogu käes , kokkuvõtted ja preemiasaajad trükivad ära järgmises Uman Lehes .
Reference	Praegu on jutud hindamiskomisjoni käes, kokkuvõte ja preemiasaajad trükitakse ära järgmises Uma Lehes.

Source	Võrokõsõ välläkutsõ mäng härgüt' kõnõlõmist pruuvma.
baseline	Võrukese iseloom ja eluviisid.
multilingual	Võrukeste väljakutses mängis härgitsesid kõnelemist proovida.
Para+BT1+FT1	Võrukeste väljakutsumine mängis innustust rääkimist proovima.
Para+BT1&2+FT1&2(*)	Võrukese väljakutsele mängib innukalt kõnelusi proovima.
Reference	Võrokese väljakutse mäng ärgitas kõnelemist proovima.

Source	Nii om võimalus telefon võita ka Uma Lehe teljäl.
baseline	Nii on võimalus telefongu ka Uma Pidoga mitmeti seotud.
multilingual	Nii on võimalus telefon võita ka Uma Lehe telgis.
Para+BT1+FT1	Nii on võimalus telefon võita ka Uma Lehe telgil.
Para+BT1&2+FT1&2(*)	Nii on võimalus telefon võita ka Uma Lehe telgil .
Reference	Nii on võimalus telefon võita ka Uma Lehe tellijal.

Table 6. Võro-Estonian

Source	Ja ühe hea külje leiab siidritegija uue nime juures veel.
baseline	Ja üte hää ütest põlvõst inemiisist lõpõ nime man vahtsõ nime man.
multilingual	Ja üte hää küle om siidritegijä vahtsõ nime man viil.
Para+BT1+FT1(*)	Ja üte hää küle löüd siidritegijä vahtsõ nime man viil.
Para+BT1+BT2(*)	Ja üte hää küle lövvüs siidritegijä vahtsõ nime man viil.
Reference	Ja üte hää küle löüd siidritegijä vahtsõ nime man viil.

Source	Uue nime väljamõtlemisel oli tähtis, et oleks selge side kohaliku kogukonnaga ja et nimi aitaks jutustada ettevõtte lugu.
baseline	Vahtsõ opimatõrjaali saamisõs oll tähtsä, et tähtsä olõs ka selge sõnumiga tõsitsit luulõtuisi.
multilingual	Vahtsõ nime vällämõtlemisel oll tähtsä, et olõsi selge side paigapäälitse kogokunnaga ja et nimi avitas kõnõlda ettevõtte lugu.
Para+BT1+FT1(*)	Vahtsõ nime vällämõtlõmisõ man oll' tähtsä, et olõssi selge side paigapäälidse kogokunnaga ja et nimi avitas ettevõttõ lugu kõnõlda.
Para+BT1+BT2(*)	Vahtsõ nime vällämõtõldõn oll' tähtsä, et olõs selge side paigapäälidse kogokunnaga ja et nimi avitas kõnõlda ettevõttõ lugu.
Reference	Vahtsõ nime vällämärkmise man oll' tähtsä, et olõs selge köüdüs paikligu kogokunnaga ja et nimi avitanuq jutustaq ettevõtmisõ luku.

Source	Samas tegutsevad kotkad kultuurmaastikul, mis tähendab, et ka inimesel on tähtis roll selles, et neil hästi läheks.
baseline	Samal aol om mi kotustõ perändüskultuurmaastikkõ, miä tähendäs, et inemisel tähtsä om tähtsä, et näil olõ-i tähtsä.
multilingual	Saman omma' kotka' kultuurmaastikul, miä tähendäs, et ka inemisel om tähtsä roll tan, et näil häste lääsi.
Para+BT1+FT1(*)	Samal aol omma kotka kultuurmaastikul, miä tähendäs, et ka inemisel om tähtsä roll tuun, et näil häste lääsi.
Para+BT1+BT2(*)	Saman toimõndasõq kotkaq kultuurmaastikul, miä tähendäs, et ka inemisel om tähtsä roll tuun, et näil häste lääsiq.
Reference	Saman toimõndasõq kotkaq kultuurmaastikul, miä tähendäs, et ka inemisel om tähtsä roll tuu man, et näil häste lännüq.

Source	Lähem info kampaania kohta kodulehelt kl.ee.
baseline	Teedüst saa mano küssü' kooriliikõst.
multilingual	Ütskõik teedüs kampaania kotsilõ kodolehe päält kl.ee.
Para+BT1+FT1(*)	Läts' teedüs kampaania kotsilõ kodolehe päält kl.ee.
Para+BT1+BT2(*)	Lähkümb teedüs kampaania kotsilõ kodolehe päält kl.ee.
Reference	Ligemb teedüs kampaania kotsilõ kodolehe kl.ee päält.

Source	Kõikide auhinnasaajatega võetakse ühendust.
baseline	Kõik huvilisõq ommaq oodõduq kullõma.
multilingual	Kõikide avvuhinna saajidõ pääle võetas ühendust.
Para+BT1+FT1(*)	Kõik' avvohinna saajidõga võetas üten.
Para+BT1+BT2(*)	Kõigilõ avvuhinnasaajilõ võetas ühendust.
Reference	Kõiki avvohinna saajidõga võetas kontakti!

Source	Leevakul elab ametlikult pea 300 inimest.
baseline	Leevälapjo eläs tähtsä pää inemist.
multilingual	Leväkul eläs virallisesti pää 300 inemist.
Para+BT1+FT1(*)	Leevakul eläs virallisesti pää 300 inemist.
Para+BT1+BT2(*)	Leevakul eläs virallisesti pia 300 inemist.
Reference	Leevakul eläs kirjo perrä pia 300 inemist.

Table 7. Estonian-Võro

Source	Uutta nimeä keksiessä oli tärkeää, että olisi selkeä yhteys paikalliseen yhteisöön ja että nimi auttaisi kertomaan yrityksen tarinan.
baseline	M uhccin muitalin lei dehálaš, ahte livččii čielga oktavuohta báikkálaš servodahkii ja ahte namma veahkehivččii muitalit lihkastagaide.
multilingual	Ođđa namma lei dehálaš, ahte livččii čielga oktavuohta báikkálaš servošii ja ahte namma veahkehivččii muitalit fitnodaga máidnasiid.
Para+BT1	Ođđa nama huksemis lei dehálaš, ahte livččii čielga oktavuohta báikkálaš servodahkii ja ahte namma veahkehivččii muitalit fitnodaga máidnasa.
Reference	Ođđa nama hutkkadettiin lea dehálaš, ahte livčče čielga oktavuohta báikkálaš servvodahkii ja ahte namma veahkehivčče muitalit fitnodaga muitalusa.

Source	Samanaikaisesti kotkat toimivat kulttuurimaisemassa, joka tarkoittaa, että myös ihmisillä on tärkeä rooli tässa, jotta heille kävisi hyvin.
baseline	Ovttaáigásaččat ruovttueatnamis doibmet kulturbirrasis, mii dárkkuha, ahte maiddái olbmos lea dehálaš rolla vuordimis, vai sidjiide šattaše bures.
multilingual	Ovttaáigásaččat kotkat doibmet kulturmálles, mii dárkkuha, ahte maiddái olbmuin lea dehálaš rolla duođas, vai sidjiide livččii hui bures.
Para+BT1	Seamma áigge kotkat doibmet kulturmáilmmis, mii dárkkuha, ahte maiddái olbmuin lea dehálaš rolla dás, vai sidjiide dáhpáhuvašii bures.
Reference	Seammás goaskimat doibmet kulturduovdagiin, mii oaivvilda dan, ahte maiddái olbmuin lea dehálaš rolla dás, vai sidjiide geavašii bures.

Source	Kansa kokoontuu entiseen koulutaloon, jossa on myös kirjasto.
baseline	Riikkabeaivevahku čoahkkana ságadoalliriikkas, mas leat maid girjerájus.
multilingual	Álbmoga čoahkkana ovddeš skuvllas, mas lea maid girjerádju.
Para+BT1	Álbmot čoahkkana ovddeš skuvladássái, mas lea maid girjerádju.
Reference	Álbmot čoahkkana boares skuvlavistái, mas lea maiddái girjerádju.

Source	Nuoria on tullut tilalle aika lailla, ehkä ottaa eräs nuori kyläelämän vetämisen haltuunsa.
baseline	Nuorat leat boahtán sadjái áigi, soađi, kántorin jos čađahat gilvun.
multilingual	Nuorat leat boahtán sadjái áiggi láđje, soaitá váldit ovtta nuorra giliellima jođiheami.
Para+BT1	Nuorat leat boahtán sadjái áige láđje, soaitá váldit ovtta nuorra giliellima jođiheami háldui.
Reference	Nuorat lea boahtán lasi oalleláhkái, gánske muhtun nuorra váldá giliellima geassima ieža háldui.

Source	Kilpailuihin lähetettiin 81 juttua, kirjoittajia oli 45.
baseline	Gilvu sáddejuvvojedje 81-87, čállin lea 45.
multilingual	Gilvvuide sáddejuvvui 81 dáhpáhusa, čállit ledje 45.
Para+BT1	Gilvvuide sáddejuvvui 81 juttua, čállit ledje 45.
Reference	Gilvvuide sáddejedje 81 čállosa, čállit ledje 45.

Source	Sosiaalisessa mediassa pitivät ihmiset eniten tehtävästä “Puhu tai postaa yksi vitsi tai tarina võron kielellä”.
baseline	Sosiála medias atne olbmot eanemus barggus “ Ominayak oktavuođaváldimiid dehe máidnumaõjjstõõllâmǩerjj lea oaivvilduvvon.
multilingual	Sosiála medias doalai olbmuid eanemus bargguin ”Puhu dahje poasta okta nja dahje máidnasa gillii.
Para+BT1	Sosiála medias dolle olbmot eanemusat bargguin “Puhu dahje poasta okta njaš dahje máidnasa vuonagillii”.
Reference	Sosiála medias liikojedje olbmot maiddái bargobihtás “Muital dahje postte ovtta cukcasa dahje máidnasa võro gillii.”

Table 8. Finnish-North Saami

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Maali Tars**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Improving translation for low-resource Finno-Ugric languages with Neural Machine Translation models,

   supervised by Andre Tättar.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Maali Tars
14/01/2021