UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Kristiina Kuningas

Training the Best Neural Machine Translation Model for the Estonian-English Language Pair

Bachelor's Thesis (9 ECTS)

Supervisor Andre Tättar, MSc

Tartu 2021

Training the Best Neural Machine Translation Model for the Estonian-English Language Pair

Abstract:
To this day, a lot of neural machine translation models have been developed to produce high-quality translations on many language directions. The same goes for Estonian-English. However, these models that have been trained on that language pair are mostly multilingual or already outdated and need enhancing. This bachelor's thesis represents a bilingual approach using recent effective technologies with the most current data available to improve the previous best result for this Estonian-English language pair. This paper introduces a state-of-the-art bilingual neural machine translation system, which outperforms the previous best result achieved for Estonian-English. The system uses different methods to achieve the goal - trains baseline models on parallel data, generates additional data with available monolingual data and backtranslation, combines the synthetic data with the initial parallel corpus, trains a new model on the augmented corpus, and in the final step, uses ensembles of those already trained models.

Keywords:
neural networks, machine translation, BLEU, language technology

CERCS:
P176 Artificial intelligence

Parima närvivõrkudel põhineva masintõlkemudeli treenimine eesti-inglise keelepaarile

Lühikokkuvõte:
Tänaseks päevaks on maailmas treenitud suurel hulgal närvivõrkudel põhinevaid masintõlkemudeleid, et pakkuda paljudele keeltele kõrge kvaliteediga tõlkeid. Seda sama on tehtud ka eesti-inglise keelepaarile. Küll aga on selle keelepaari jaoks treenitud mudelid enamasti mitmekeelsed või vananenud ning vajavad täiustamist. See bakalaureusetöö esitleb kahekeelset lähenemist kasutades uusi efektiivseid tehnoloogiaid ja kõige värskemaid saadaolevaid andmestikke selleks, et parendada seni parimat tulemust eesti-inglise keelepaaril. Selles töös treenitakse eesti keelest inglise keelde parimat tõlkekvaliteeti pakkuv masintõlkemudel, mis ületab ka varasema parima tulemuse. Soovitud eesmärgi saavutamiseks kasutatakse erinevaid meetodeid - treenitakse paralleelsetel andmetel põhinev baasmudel, sünteesitakse saadaolevate monokeelsete andmete ja tagasitõlke abil uus andmestik, kombineeritakse esialgne paralleelandmestik uue sünteetilise andmestikuga, treenitakse mudel suurenenud paralleelkorpusel ning lõpuks komplekteeritakse need juba treenitud mudelid kokku.

Võtmesõnad:

tehisnärvivõrgud, masintõlge, BLEU, loomuliku keele töötlus

CERCS:
P176 Tehisintellekt

Contents

1 Introduction

Neural machine translation (NMT) is one solution for generating texts from one language to another. These machine-generated texts are used by lots of digital systems or applications, by people who just want to communicate or work daily as translators. In addition, several research groups work with these every day. Therefore it is important to continuously improve the quality of these translations.

NMT models are trained on high- and low-resourced as well as medium-resourced languages. One of the medium-resourced language pairs is Estonian-English and many models have been trained on this language pair as well. However, these models are sometimes outdated and mostly multilingual. This thesis, however, develops a bilingual system, uses most current technologies that have been recently approved by others and the latest available data. The goal with this system is to produce translations, which quality outperforms the best result so far and thereby see, how different approaches work in practice.

This bachelor's thesis aims to train a state-of-the-art machine translation model, which generates high-quality translations for Estonian-English. The quality of the translation will be evaluated and compared with the previous best result with the help of the Bilingual Evaluation Understudy (BLEU) score [1]. To the knowledge of this thesis author, the previous state-of-the-art system on Estonian-English language pair was achieved in 2018 by Tilde[1] and is introduced in the paper by Pinnis et al [17].

To fulfil the goals of this paper, three different approaches are implemented. At first, two baseline models are trained on prepared and preprocessed parallel corpus. Secondly, Edunov et al. [7] have confirmed that backtranslation can raise the quality of translation. Additional synthetic corpora can be generated with the help of available monolingual data and aforementioned backtranslation. Therefore, this approach will be used to produce new synthetic parallel data. This new corpus will be used for training the final models. In addition, relying on previous effective experiences with ensembling [7, 10, 26], this thesis is also using ensembling to improve the results even more.

This paper is divided into six bigger sections. Introduction part introduces and explains the thesis aim and gives an overview of the structure. Secondly, there is a section for introducing the background of machine translation and thesis-related terminology to give the reader a better understanding of the upcoming paper. Third, related works section introduces three papers, which are directly related to the thesis and which are the base for reaching the goal of this work. The next part - methodology - goes more into detail about the data preparation, training and evaluation. At last, there is a results section to analyse the results of trained models both quantitatively and qualitatively. The work ends with a conclusion of the thesis and discusses some possible ideas for experiments in the future.

[1] https://www.tilde.com/

2 Background of Machine Translation

2.1 Transformers

Over the last decade, statistical machine translation (SMT) has been completely overtaken by neural machine translation (NMT). In 2017 Vaswani et al. [3] introduced the new state-of-the-art NMT architecture called Transformer, which produces highest results in machine translation (see Figure 1).

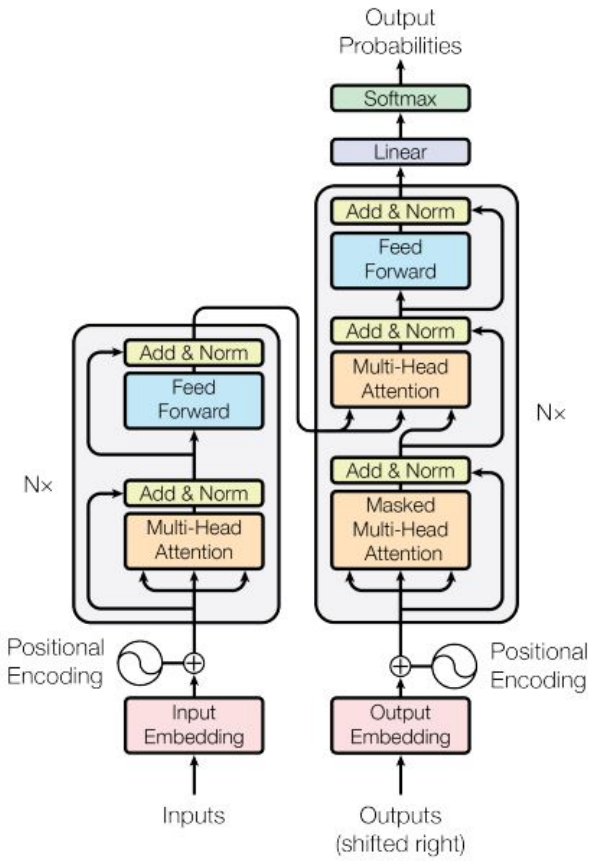

Figure 1. Transformer architecture [3].

Similarly to previous Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), Transformers are based on encoder-decoder architecture (see Figure 1) [3], but have left behind the recurrence and convolution [6]. RNNs use sequential data processing, meaning tokens in sequences are processed one after another [6]. Because of this sequential data processing and memory limit, it is also hard not to say impossible to process longer sequences [6]. In addition, it takes up a lot of time to train [3]. CNNs and

Transformers do not have to prosecc sequences one after another, unlike RNNs. CNNs process data hierarchically and therefore allow some parallelization, however, they take up a lot of memory [6]. Transformers also allow parallelization, but they have a whole different approach. They rely entirely on self-attention and positional-embedding [3], which eliminate the issues of previous methods and have been proved to perform better as well.

According to Vaswani et al. [3], at the very beginning, the input sequence is embedded to vectors. Embedded vectors are then positionally encoded because, unlike RNNs, Transformers' input sequence is not inserted in sequentially and Transformers do not know the word order in the sequence. Therefore positional encoding of the embedded word vectors is necessary in order to get to know positional information about tokens in the sequence. Then these aforementioned word vectors that are already positionally encoded, are sent to the encoder layer, which consists of a multi-head attention layer, a feed-forward network and normalisation following them both [3] (see Figure 1).

Then this encoder output is turned into attention vectors, which are sent to the decoding layer. Decoding layer has a masked multi-head attention layer and a feed-forward network, but also a multi-head attention layer in between. Then, every token of the output sequence is iteratively generated with every decoding step until the specific character representing the end of the output sequence has been reached. In the end, with the help of softmax and linear layer, these vectors are turned back into a word [3].

The main virtue of Transformers is self-attention [3], which was represented in several layers of the architecture described before. Tang et al. [6] said in their work that self-attention helps to connect tokens in a sequence to another token in the same sequence. This helps the encoder to look at and focus on different tokens at the same time and therefore produce a better and more precise output more directly related to the context [6].

While with RNN, the distances between tokens got larger after every additional token, then with self-attention the distances remain the same for all connections [6]. This eliminates the possibility of forgetting about tokens further away. What is more, Transformers do use multi-head attention instead of just single head attention according to Vaswani et al. [3]. It provides a more fine-grained version of self-attention [6].

2.2 BLEU score

The evaluation of machine translation is inevitable and necessary not only to measure the effectiveness of translation but also to improve the quality of it in time. The principle idea behind machine translation evaluation is to compare the machine translation with a reference sentence translated by humans. An extensive, but at the same time very expensive method for evaluation is manual human translation evaluation. It is the most precise and accurate solution, however, it could take weeks or even months to finish [1].

That is where automatic evaluation metrics come under discussion as they calculate the evaluation scores automatically without the time cost.

Bilingual Evaluation Understudy (BLEU) is the dominant metric for automatic machine translation evaluation as it provides fast and language-independent evaluation [1]. Additionally, it has a high correlation with human evaluation. BLEU's idea for evaluation is to compare *n*-grams of machine-translated sentences to reference sentences and count all of the matches between those.

BLEU is based on the basic precision measure, however the downside of it is that machine translation systems can generate sentences with overgenerated words and therefore the outcome might seem precise, but in reality, has flaws. To eliminate this issue, BLEU uses modified *n*-gram precision, meaning if there is a match between the candidate word and the corresponding word in the reference sentence, then the reference word can not be used again [1].

Table 1 delivers an example to showcase the difference between basic and modified *n*-gram precision [1]. Modified unigram precision, in this case, would be 2/7 as two candidate sentence words match two reference sentence words and after the match is found these words are exhausted. However, the standard unigram precision would be 7/7 as it uses repeatedly the same words. It is clear that the candidate translation is not identical to the reference sentence.

Candidate Translation	the the the the the the the
Reference Sentence	The cat is on the mat.

Table 1. The example of candidate translation and reference sentence [1]. Underlined words represent the matches between sentences.

With BLEU, it is also necessary to take the *brevity penalty* factor into account, because unrealistically high scores could appear in cases where machine-translated sentences are shorter than reference sentences. For example when the reference sentence is "It is a cute dog" and the candidate sentence is "It is" then without *brevity penalty* it would be a high-level translation as all of the words in candidate translation also appear in reference sentence, however in practice it is not. To calculate *brevity penalty* score (see Equation 1) we need c, and r, the lengths of the candidate and reference sentence accordingly [1]:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{1}$$

BLEU is a product of *brevity penalty* factor and the "geometric average of the modified *n*-gram precisions, p_n, using *n*-grams up to length *N* and positive weights w_n

summing to one" [1]. See Equation 2.

$$BLEU = BP \cdot exp(\sum_{i=n}^{N} w_n \log p_n) \tag{2}$$

BLEU score ranges from 0 to 1. If the score is 0, there is no connection between the candidate and reference sentence and when it is 1, the translations are identical [1].

2.3 Backtranslation

Backtranslation is the new standard approach in producing high-quality neural machine translation as marked by Poncelas et al. [9]. Sennrich et al. [8] have introduced backtranslation as an approach for producing more training data with monolingual data and improve the model's fluency. Figure 2 represents the process of backtranslation. As stated by Edunov et al. [7], to use target monolingual data for producing the synthetic parallel data in the end, a *target-to-source* system has to be trained at first.

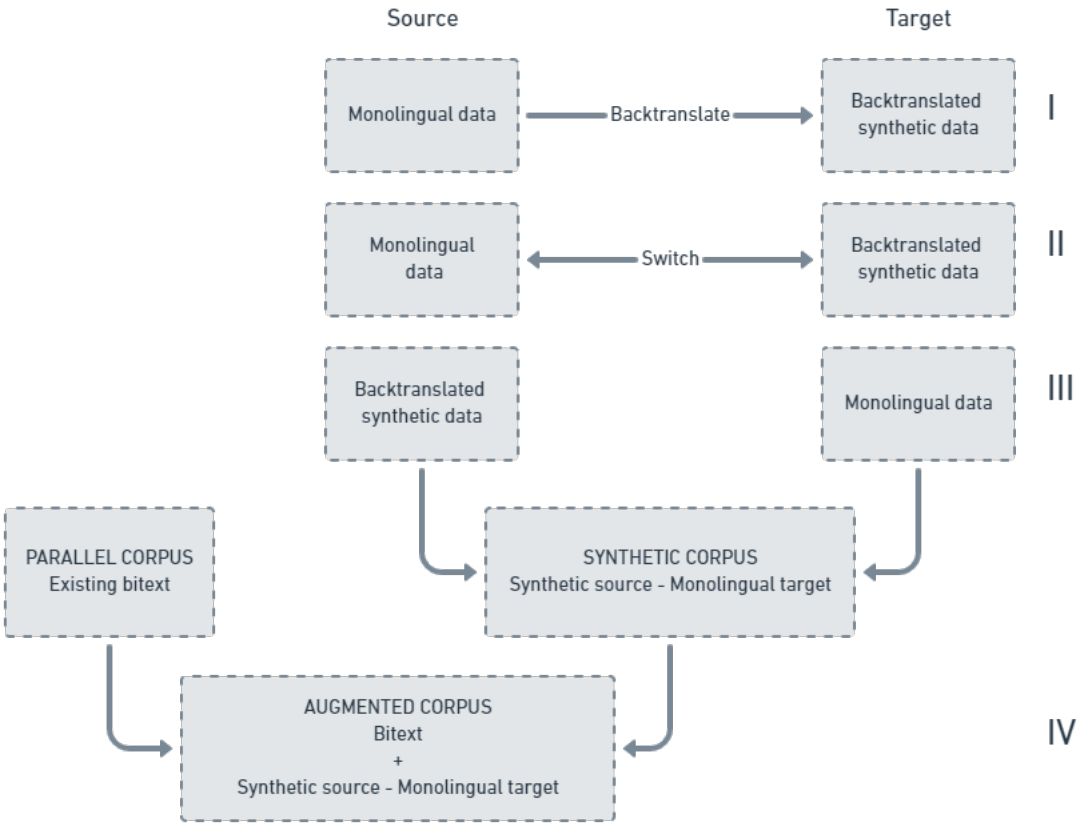

Figure 2. The process of backtranslation. Roman numerals I-IV represent the order of the steps.

The baseline *target-to-source* model is then used for translating the target monolingual data available [7] (see Figure 2). Afterwards, the backtranslation's source-side which was monolingual data and machine-translated target side are switched, meaning translated synthetic target side becomes source of the new synthetic parallel corpus and the monolingual text created by humans becomes target again [8]. The newly created

corpus is then combined with existing bitext (see Figure 2) to create a larger parallel data set to train the final system on [7].

2.4 Ensembling

Thomas G. Dietterich [20] introduces ensembles as sets of individual classifiers. They all have their individual predicitions, but the purpose of the ensemble is to combine all these predictions and make one final prediction for the whole ensemble. In his work he explains why using sets of classifiers outperforms using each classifier alone [20]. In machine translation context it means that if we combine together several backtranslation models or checkpoints, then we create an ensemble, which works better than any of the models separately.

There are a lot of algorithms based on which the ensembles work and make their final decision. The most basic method for the decision-making is based on weights, meaning the decisions of different classifiers are all taken into account, but some classifier's "vote" weighs more [20]. In addition there are also voting classification algorithms like boosting and bagging, which in fact are mainly used for ensembling. for making the final decision for the ensemble [21]. With boosting, classifiers are sent to the ensemble one after another and every model that comes after another corrects the errors of the previous model [21]. For bagging, the original data is split into samples and each of the classifiers learns from different sample. After that, the classifiers are joint and the final decision is made. Unlike boosting, classifiers work at the same time, not sequentially [21].

2.5 Byte Pair Encoding

Machine translation is an open-vocabulary problem, but in practice, neural models mainly use fixed vocabulary. Sennrich et al. [4] prove that using subword units to encode words solves the open-vocabulary challenge. Subword models help to compress the data because many words can have similar parts in them, for example, "homeless", "homeowner", "hometown" etc. and when splitting them into smaller pieces, then instead of having a separate word for everything we can produce words by combining the small pieces. In addition, splitting the words into subunits helps to produce some rare words with segments that are not included in the initial vocabulary and that are not seen at training time [4].

One algorithm to segment the words and compress the data is Byte Pair Encoding (BPE) algorithm [5]. The algorithm of BPE finds the most frequently occurring bytes that are next to each other and replaces these bytes with a new byte pair that did not exist in the original data before. The algorithm works iteratively until there exists no more frequent byte pairs or no more unused bytes that could make up a new byte pair [5]. In NMT aforementioned bytes are replaced with characters and sequences of characters [4] and by the end, an abbreviated text of subword units is produced.

For example, let us consider an example where the number represents the frequency of the word that is split into characters [4]:

’l o w </w>’: 5, ’l o w e r </w>’: 2, ’n e w e s t </w>’: 6, ’w i d e s t </w>’: 3

The most frequent characters next to each other in this example are ’e’ and ’s’ as they appear nine times in total. Following the BPE algorithm, a new character pair of ’es’ is created which at the same time replaces the standalone characters ’e’ and ’s’. The output of this step is following:

’l o w </w>’: 5, ’l o w e r </w>’: 2, ’n e w **es** t </w>’: 6, ’w i d **es** t </w>’: 3

The process is repeated until the condition for stopping the algorithm is met, meaning when there is no frequent byte pairs or unused bytes available.

3 Related Work

To fulfil the goals of this thesis, two additional approaches besides training baseline system are taken into account. This work uses backtranslation [7] and ensembling [14] to get higher results than the top so far.

3.1 Backtranslation and Synthetic Data

The previous best-performing translation systems on Estonian-English language pair were developed in 2018 when Tilde[2] submitted its NMT systems to a shared task in the Third Conference on Machine Translation (WMT 2018) [17]. Previously in 2017, they had competed with multiplicative long short-term memory (MLSTM) units and in 2018 they showed that Transformer models outperform previous MLSTM models by approximately 4 BLEU points in both language directions. [17]

Pinnis et al. [17] submitted altogether seven different constrained and unconstrained Transformer models. Data used for constrained models was given by 2018 WMT organisers, however, for unconstrained models they also used other publicly available data. This thesis uses publicly available data as well for training the models, however the sources of data are different from Tilde's work source.

Tilde's best-resulting system was an unconstrained system, which used backtranslation, but the backtranslated data for this system was produced by pre-existing unconstrained MLSTM-based NMT systems [17]. Contrary to Tilde's work this thesis generates additional synthetic data with baseline systems that were of Transformer architecture and trained during the thesis.

In WMT 2018 Edunov et al. [7] achieved a new state-of-the-art result on English to German translation. With their setup, they prove how synthetic data produced with backtranslation at a large scale can improve the translation results. In their case, BLEU was raised by about 2.6 points compared to baseline score. What is more, in comparison with a lot of previous approaches of using backtranslation, Edunov et al. [7] emphasise the fact that sampling or noisy data brings better results than traditionally used beam or greedy search for synthetic data generation. Imamura et al. [12] also verify the effectiveness of sampling. Briefly, sampling multiplies the source sentences that are back-translated from the target monolingual data. It helps to average the errors that come with synthetic sentences and also gives more diversity than synthetic translations usually have to be more like human translations [12].

This thesis is largely based on the related work by Edunov et al. [7]. It follows their approach, similarly uses publicly available data, uses a large scale of monolingual data, trains baseline systems for backtranslation, backtranslates data to produce synthetic data and also uses sampling in backtranslation to enhance the results even more.

[2]https://www.tilde.com/

3.2 Ensembling

OPPO's submission to the Fifth Conference on Machine Translation by Shi et al. [10] proves that ensemble models improve the translation quality. They trained several models on many different language pairs. Results showed that additional ensembling of previously trained models improved the quality of translation for all the models. In average, the BLEU score improvements after ensembling were between 0.3 and 1.0 BLEU points. For example, for English-Japanese language pair ensembling raised 0.6 BLEU points on English to Japanese and 0.7 BLEU points for Japanese-English direction [10]. Unfortunately, more detailed information, for example about the sizes of ensembles or which approach for decision-making inside the ensemble was used, was not introduced.

Experiments by Edunov et al. [7] confirm as well that ensembling can improve the quality of the translation. They ensembled six different back-translation models in their work. Compared to their baseline model, ensembling did raise the BLEU with additional 0.6 points on English to German direction compared to the previously trained backtranslation model. In addition, Pinnis et al. [17] brought out in their work that ensembling improved the translation quality for constrained models as the results with ensembles of averaged models were higher than the results of averaged models on their own. They used ensembles of 3 averaged models [17].

4 Methodology

The data preparation, model training and translation generation scripts used in this thesis are based on the paper by Edunov et al. [7] and are more or less modified compared to their approach. For preprocessing, training the models and translation generating, Fairseq[3] was being used. Fairseq is a toolkit for researchers to train "custom models for translation, summarization, and other text generation tasks" [13].

4.1 Data

As this paper is based on bilingual models only, then accordingly data of just two languages - Estonian and English - are used. Data is separated into parallel and monolingual data. Parallel data alone is used for training the baseline model, however, parallel corpus together with monolingual data is needed for backtranslation.

4.1.1 Parallel Data

All of the parallel data for training was downloaded from OPUS[4], which is a freely available corpus of parallel data [11, 16]. Data was downloaded from 34 different corpora and later combined into one large corpus consisting of 37.52 million sentence pairs. All of the details for downloaded training datasets are represented in Table 14 in the appendix. As for the test and development data, this thesis uses *newstest2018* test set and *newsdev18* development set which were provided for WMT 2018[5] and contained 2000 sentence pairs each (see Table 2).

Before training, data had to be prepared and preprocessed. At first, all of the sentences longer than 250 words and with a source-target ratio over 1:5 were removed. Data preparation consisted of punctuation normalisation and removing nonprintable characters. Furthermore, parallel data was tokenised with Moses, however, it was not necessary and later in monolingual data preparation this step was skipped. All of these processes were done with Moses[6] scripts.

During the data preparation, leaky data was found, meaning that test and valid data also occurred in training data. Overlap like this will make the results too optimistic as the model already trains on the data that is meant for testing. Therefore, data was cleaned of overlap to avoid results distortion later on. The amount of data that remained after data preparation and overlap elimination can be seen in Table 2.

[3] https://github.com/pytorch/fairseq
[4] https://opus.nlpl.eu/
[5] http://www.statmt.org/wmt18/translation-task.html
[6] https://github.com/moses-smt/mosesdecoder

After preparing the data with Moses and removing leaky data, BPE algorithm was implemented. This was done with SentencePiece[7] described by Kudo and Richardson [15], which tokenises and detokenises text and unlike Moses, SentencePiece is language independent. SentencePiece uses training data to learn where to split words and then applies the same model on training, test and development data. The vocabulary size for BPE was 32 000 and the character coverage was 0.9995. In Tilde's work [17] vocabulary size was 50 000 for their best model.

The final step before training was preprocessing and this thesis used the help of the *fairseq-preprocess* tool for it. Preprocessing helps to join the vocabulary of source and target data and binarizes data for training.

	Training data	**Test data**	**Development data**
Before preparation	37 522 828	2000	2000
After preparation	37 488 564	2000	2000

Table 2. The number of sentence pairs of parallel training, development and test data before and after data preparation.

4.1.2 Monolingual Data

Monolingual data was needed for backtranslation and it was downloaded from the WMT 2018 submission task's web page[8]. At first downloaded sets consisted of 20 974 127 Estonian and 26 861 181 English sentences as can be seen in Table 3. After downloading the data, 20 million sentences of both languages were picked for data preparation. Preparation steps were the same as for parallel data - punctuation normalisation and removing nonprintable characters. In addition, duplicated sentences in Estonian monolingual data were removed. For English, it was not necessary as downloaded data was already deduplicated.

In the end, after the preparation steps, data was split into shards. Each of the shards included one million sentences, except the last one. After monolingual data preparation and splitting into shards, 19.99 million English and 18.45 million Estonian sentences were used for backtranslation and synthetic data generation (see Table 3). Finally, the subword segmentation and preprocessing part was the same as for parallel data and done with the help of SentencePiece and *fairseq-preprocess*. Later on, while combining the backtranslated data with parallel data, sentences containing over 250 words and the ones that had a source-target ratio over 1:5 were removed.

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

[8]http://www.statmt.org/wmt18/translation-task.html

Language	Before preparation	After preparation
Estonian	20 974 127	18 454 844
English	26 861 181	19 999 165

Table 3. The number of sentence pairs of monolingual data before and after data preparation.

4.2 Training

4.2.1 Settings and Environment

Fairseq-train command -line tool was used for training all of the models. The main hyperparameters remain the same for all of the training processes, apart from the source and target language and source-destination data folder parameters.

As a lot of data was involved in the training process, then the training took place at the University of Tartu High Performance Computing (HPC) Center[9] and 1 GPU was being used for all of the trainings. Models trained for approximately 7 or maximum 8 days.

All of the models used Transformer architecture [3] with preset structure. Related works used also Transformer architecture for their best-performing models. Although, Edunov et al. [7] used Big Transformer architecture, not base Transformer architecture that is used in this thesis to reduce training time. Models had 6 encoding and 6 decoding layers and the embedding dimensionality of produced vectors was 512. 8 parallel attention layers were used, so the dimension of each head was 64. Pinnis et al. [17] used an embedding size of 512 as well for its most resultative model, however, their model had 7 encoder and decoder layers.

Models used Adam optimizer [14] with following parameters: β_1 = 0.9 and β_2 = 0.98. For the learning rate, inverse square-root scheduler [3] was used with the learning rate of 0.001 and 4000 warm-up steps, meaning learning rate linearly increased until warm-up training steps were reached and then decreased to the "inverse square-root of the step number". Dropout and weight decay are for regularisation. Dropout probability p-value was 0.3 and weight decay was 0.0. Label smoothing cross-entropy [18] was used with label smoothing parameter value 0.1. The maximum number of tokens in a batch were 15 000. After every 5000 updates model was saved to a checkpoint and the last 32 checkpoints were saved. Update frequency was 8.

4.2.2 Baseline and Backtranslation Models

During the thesis, two Transformer baseline models were trained - one from Estonian to English and the other *vice versa*. Only an initial parallel corpus was needed for training

[9] https://hpc.ut.ee/en/home/

them. Baseline systems are usually requisite for analysing the enhancement afterwards and also necessary for the backtranslation.

In virtue of papers that have verified the benefits of backtranslation [2, 10], the approach was utilised in this work also. In the synthetic data developing process, sampling was used instead of the beam or greedy search to strengthen training signal of the data and therefore improve the results even more. Edunov et al. [7] had previously proven its efficiency as their BLEU score improved with and especially with a larger amount of synthetic data.

During this work, only one cycle of backtranslation was run. To generate a combined corpus of parallel and synthetic data, for example on Estonian to English direction, previously prepared English monolingual data was translated to Estonian. Then the source sentences and target synthetic data were switched, so monolingual reference sentences in English were now target and machine-translated Estonian sentences were source. Generated new parallel corpus was then added to the baseline parallel corpus.

This combination of parallel and synthetic corpus was needed for training a new model, which now has more data to train on. The same process was done for English to Estonian direction with Estonian monolingual data.

4.2.3 Ensembling

As proved by previously introduced works by Shi et al. [10] and Edunov et al. [7], ensembling improves the quality of translation. In this thesis, both baseline and backtranslation models were trained on the same data, but in the end, they all perform differently. Every checkpoint of the model performs differently. With ensembling, we gather some of these models and checkpoints together and combine them into an ensemble, which is then used for generating the translations.

To enhance the BLEU in this thesis, different variations of ensembles were tried and the best-resulting one was taken into account for the final result. The process of ensembling takes place in the translation generation process. Ensembles were comprised of different amounts of checkpoints of baseline and backtranslation models. Ensembles of two, four, or six checkpoints of the same or different models were generated and used for translation. For instance, there were ensembles of only two backtranslation model checkpoints, but there were ensembles that were a combination of one baseline and one backtranslation model checkpoint as well.

4.3 Evaluation

The evaluation metric used to judge the quality of translations was BLEU. Videlicet, an implementation of BLEU called SacreBLEU[10] was utilised [2]. While BLEU values varied from 0 to 1, SacreBLEU can be a number between 0 and 100.

This thesis used Fairseq command-line tools like *fairseq-generate* and *fairseq-interactive* to translate the WMT 2018 test sets that were automatically downloaded by SacreBLEU and tokenized with SentencePiece. Before using SacreBLEU, the translated text had to be detokenised and Sentencepiece had to be removed. In the end, the SacreBLEU score is produced by comparing the reference sentences with generated translations and this score could be then used to compare the results of different models using the same test set [2].

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

5 Results

Quantitative as well as qualitative analysis of the results was done. With quantitative analysis, the comparison between earlier best results by Tilde [17] could be done to understand if the thesis has fulfilled its goals. Qualitative analysis subjectively discusses the improvements between reference and machine-translated sentences.

5.1 Quantitative Analysis

Quantitative analysis was based on the evaluation process utilising BLEU with SacreBLEU using WMT 2018 test set. The BLEU scores are represented in Table 4. The best acting models were picked for the final results, meaning several checkpoints were tested out to get the highest result. In addition, one of the models is an ensemble and the best combination of checkpoints is represented in Table 4.

The best baseline model from Estonian to English direction achieved 30.9 BLEU points. Adding backtranslation to the baseline raised the scores by 0.4 BLEU points. The best model, the ensemble of three baseline model checkpoints and three backtranslation model checkpoints, gained BLEU by 0.8 points compared to the baseline.

For comparing the results with Tilde's previous work and getting the most accurate results, Tilde's translations on the same test set were downloaded and evaluated with the same SacreBLEU signature than all the other models in this thesis. The results did show that Tilde's state-of-the-art result was 30.9 BLEU points and it was also marked in the paper itself [17]. This thesis' baseline model achieved the same score as their best-resulting one. As backtranslation and ensembling scores improved the results, then in comparison with Tilde's best result the best model of this system gives an outcome of 0.8 points higher BLEU score (see Table 4).

Model	**Estonian-English**	**English-Estonian**
Tilde WMT18 [17]	30.9	
Baseline	30.9 / -	22.2 / -
Baseline + BT	31.3 / +0.4	22.3 / +0.1
Baseline + BT + Ensemble	31.7 / +0.8	22.7 / +0.5

Table 4. BLEU scores and improvements compared to baseline scores. *Tilde WMT18* - state-of-the-art system by Tilde from WMT18, *Baseline* - model trained on parallel data, *Baseline + BT* - model trained on parallel + backtranslated data, *Baseline + BT + Ensemble* - ensemble of baseline and backtranslation models.

Translations in the opposite direction had also improvements after implementing backtranslation and ensembling, however, the results were not higher than in Tilde's

work [17]. Backtranslation helped to gain 0.1 BLEU points compared to the baseline score of 22.2 BLEU points (see Table 4). Ensembling scores were 0.5 BLEU points higher than the baseline score.

Quantitative results for ensembles did show that the best results for both language directions were achieved with six checkpoints (see Table 5). Overall, the bigger the ensemble was, the better the results got. Ensembles with two checkpoints had BLEU results of 31.5 and 22.5 for Estonian-English and English-Estonian accordingly. Every additional two checkpoints did raise the score by 0.1 BLEU points. The differences were not remarkable, but positive nevertheless.

Ensemble	**Estonian-English**	**English-Estonian**
Two checkpoints	31.5 / -	22.5 / -
Four checkpoints	31.6 / + 0.1	22.6 / 0.1
Six checkpoints	31.7 / + 0.2	22.7 / + 0.2

Table 5. BLEU scores and improvements compared to the smallest ensemble. All of the ensembles were the best-resulting combinations of baseline and backtranslation model checkpoints.

The ensembles that were a combination of baseline model and backtranslation model checkpoints performed better than the ensembles which included the checkpoints of one model only. This could be because of the additional diversity that comes with different models. Hansen and Salamon [19] have said that the ensembles are more accurate, when the classifiers that create the ensemble are more diverse. For example, for English-Estonian direction, the best-performing ensemble of two backtranslation model checkpoints had a score of 22.2 BLEU points, however, a combination of baseline and backtranslation models had a BLEU score of 22.5. All of the other models had as well approximately the same differences when comparing the same size combination ensembles with one model ensembles only.

In conclusion, relying on the outcome seen in Table 4, results did improve after every approach that was taken into account for enhancement. The main goal of getting a high score for Estonian to English direction translation was successfully attained. Regardless of not getting a higher score than the state-of-the-art system has for English to Estonian translation, the theory that was implemented to improve the quality throughout the process worked in practice. In addition, the results in Table 4 showed that the translation quality got better when the ensemble contained more checkpoints.

5.2 Qualitative Analysis

The qualitative analysis represents how machine-translated texts differ from the reference sentences. In addition, it shows changes and possible improvements between translations of different models. All of the source and reference sentences in the examples are taken from WMT 2018 test set. Furthermore, translations generated by the University of Tartu Natural Language Processing Group's model[11] (UT NLP) are added to the analysis.

5.2.1 Estonian-English

In the first example (See Table 6) we can see a change and progress after every translation and none of the translations are identical to the reference sentence. Despite the changes in the words, the meanings of the sentences remain the same. The translations of all the models seem to be better than the reference sentence itself, at least for the first part of the sentence, as they give a more correct and clear translation than it is for the reference sentence. "they were released the same day" seems grammatically more correct and more logical than saying "the same day they were released". Baseline and ensemble models' translations are the closest to the reference sentence as "there's an investigation now" and "there's an investigation going on" represent the meaning of "while an investigation is currently underway" most precisely. UT NLP translation was identical to the ensemble model.

[11] https://neurotolge.ee/

Source	Nad lasti samal päeval vabaks, aga praegu toimub uurimine.
Reference	The same day they were released, while an investigation is currently underway.
Baseline	They were released the same day, but there's an investigation now.
Baseline + BT	They were released the same day, but they're under investigation.
Baseline + BT + Ensemble	They were released the same day, but there's an investigation going on.
UT NLP	They were released the same day, but there's an investigation going on.

Table 6. The first example sentence on Estonian-English language direction with reference sentence and corresponding translations. The most important differences between translations are highlighted.

Table 7 represents the qualitative translation results of the second source sentence. Although the words used for the reference sentence and the best model translation differ in some way, the meaning remains generally the same. "Increased restrictions" in the machine-translated sentences are more directly associated with Estonian *suurendanud vaba telkimise piiranguid* than reference sentence's approach "tightened control over free camping". Translation of the best model differs from the baseline model only by that the word *harsh*, which becomes *severe*. UT NLP model adds the article "the" in front of "restrictions" and also matches with the reference sentence by using "is caught" instead of "gets caught", which was the case for thesis' models.

Source	Kreeka ametivõimud on suurendanud vaba telkimise piiranguid ja kehtestavad juba praegu karme trahve kõigile, kes jäävad vahele puhkamisega kohas, mis pole selleks ette nähtud.
Reference	The Greek authorities have tightened control over free camping and are already imposing serious fines on anyone who is caught resting in a place not intended for the purpose.
Baseline	The Greek authorities have increased restrictions on free camping and are already imposing harsh fines on all those who get caught on holiday in a place that is not intended for this.
Baseline + BT	The Greek authorities have increased restrictions on free camping and are already imposing severe fines on anyone who gets caught resting in a place not intended for that purpose.
Baseline + BT + Ensemble	The Greek authorities have increased restrictions on free camping and are already imposing severe fines on anyone who gets caught on holiday in a place that is not intended for this.
UT NLP	The Greek authorities have increased the restrictions on free camping and are already imposing severe fines on anyone who is caught resting in a place .

Table 7. The second example sentence on Estonian-English language direction with reference sentence and corresponding translations. The most important differences between translations are highlighted.

As we can tell by the sight of Table 8, translations generated by thesis models remain the same but differ from the reference sentence and also the UT NLP model's translation.

Differences between the reference sentence and all the translated sentences bear on the verbs "detaining" and "arresting", which in theory mean the same. In UT NLP translation we can also see that unlike all the other sentences "In Greece" is put in the beginning of the sentence instead of the end. The similarity between UT NLP and reference sentence that other models do not have is that "arrested" is in the past tense, but "were arrested" or "were detained" is the past participle.

Source	Kreekas vahistati ebaseadusliku telkimise eest kaks bulgaarlast.
Reference	Two Bulgarians were detained for illegal camping in Greece.
Baseline	Two Bulgarians arrested for illegal camping in Greece.
Baseline + BT	Two Bulgarians arrested for illegal camping in Greece.
Baseline + BT + Ensemble	Two Bulgarians arrested for illegal camping in Greece.
UT NLP	In Greece, two Bulgarians were arrested for illegal camping.

Table 8. The third example sentence on Estonian-English language direction with reference sentence and corresponding translations. The most important differences between translations are highlighted.

Table 9, on the other hand, shows an example where machine-translated texts did not improve the translation and produced incorrect translations. Already the baseline model's translation is false and the other three translations including UT NLP's translation change the sentence formulation a little, but do not produce a better translation.

Source	Sõdureid, jah, liigub meil palju!
Reference	Yes, we have a lot of soldiers about town!
Baseline	Soldiers, yes, we have a lot of moves!
Baseline + BT	Soldiers, yes, we move a lot!
Baseline + BT + Ensemble	Soldiers, yes, we've got a lot of moves!
UT NLP	Soldiers, yes, we're moving a lot!

Table 9. The fourth example sentence on Estonian-English language direction with reference sentence and corresponding translations. The most important differences between translations are highlighted.

All in all, these translations in the examples show that the models may differ a little from each other and reference sentence but nevertheless offer a coherent translation. The translations of UT NLP model were very similar to the thesis ones. The biggest difference between thesis models' and UT NLP model translations was with the longest sentence (see Table 7). On the other hand, as Table 9 did show, the models can produce incorrect translations as well, because they are not perfect and have flaws that can still be inexplicable. However, we can not make any fundamental conclusions on the whole translation quality with these four examples analysed.

5.2.2 English-Estonian

In the first example for the English-Estonian language (see Table 10), the biggest difference between reference sentence and translations is the phrase order. The reference sentence outperforms the translated outputs just by the word and phrase order. The reference sentence sounds more natural and correct, at least for the first part of the sentence. Although the translations are not high quality, they still provide an understandable meaning.

Source	The same day they were released, while an investigation is currently underway.
Reference	Nad lasti samal päeval vabaks, aga praegu toimub uurimine.
Baseline	Samal päeval nad vabastati, samal ajal kui uurimine on käimas.
Baseline + BT	Samal päeval nad vabastati, samal ajal kui praegu on käimas uurimine.
Baseline + BT + Ensemble	Samal päeval nad vabastati, samal ajal kui uurimine on käimas.
UT NLP	Samal päeval nad vabastati, samal ajal kui uurimine on käimas.

Table 10. The first example sentence on English-Estonian language direction with reference sentence and corresponding translations. The most important differences between translations are highlighted.

Table 11 represents the translations of a longer Estonian sentence. In this example, the best translation is provided by the best model, the ensemble model. Backtranslation model was the only one, who translated the word "free" to *tasuta*, meaning free of charge, not *vaba*, which has the freedom-related meaning. The phrase "who is caught resting" is translated most correctly by ensemble model and the other models, including UT NLP models, provide a translation such as "püütud puhata kohas" or "püütud puhkama

kohas" or "püütud puhkama kohta", which could seem like a correct word-for-word translation, but actually is not. UT NLP translation differs from thesis model translations and reference sentence by using past tense instead of present and by the meaning of the Estonian translation of the phrase "have tightened control over free camping".

Source	The Greek authorities have tightened control over free camping and are already imposing serious fines on anyone who is caught resting in a place not intended for the purpose.
Reference	Kreeka ametivõimud on suurendanud vaba telkimise piiranguid ja kehtestavad juba praegu karme trahve kõigile, kes jäävad vahele puhkamisega kohas, mis pole selleks ette nähtud.
Baseline	Kreeka ametiasutused on karmistanud kontrolli vaba telkimise üle ja kehtestavad juba tõsiseid trahve kõigile, kes on püütud puhkama kohas, mis ei ole selleks otstarbeks ette nähtud.
Baseline + BT	Kreeka ametiasutused on karmistanud kontrolli tasuta telkimise üle ja määravad juba tõsiseid trahve kõigile, kes on püütud puhata kohas, mis ei ole ette nähtud selleks otstarbeks.
Baseline + BT + Ensemble	Kreeka ametiasutused on karmistanud kontrolli vaba telkimise üle ja juba kehtestavad tõsiseid trahve kõigile, kes jäävad puhkama kohas, mis ei ole selleks otstarbeks ette nähtud.
UT NLP	Kreeka ametiasutused on karmistanud vabade telkimiskohtade kontrolli ja on juba kehtestanud tõsiseid trahve kõigile, kes on kinni püütud puhkama kohta, mis ei ole selleks ette nähtud.

Table 11. The second example sentence on English-Estonian language direction with reference sentence and corresponding translations. The most important differences between translations are highlighted.

The third example (see Table 12) shows again how sentences can differ by word or phrase order. Only backtranslation model's output, as well as reference sentence, start the sentence with *Kreekas*, in all other sentences "in Greece" is put in the middle. The

word "camping" has been translated as *matkamine* by backtranslation model and *laagrid* by UT NLP model, but not as *telkimine* which is the reference translation. In addition, compared to the reference sentence, which uses *vahistati* for the translation of "detained", translations use verb *peeti kinni*.

Source	Two Bulgarians were detained for illegal camping in Greece.
Reference	Kreekas vahistati ebaseadusliku telkimise eest kaks bulgaarlast.
Baseline	Kaks bulgaarlast peeti Kreekas ebaseadusliku telkimise eest kinni.
Baseline + BT	Kreekas peeti ebaseadusliku matkamise eest kinni kaks bulgaarlast.
Baseline + BT + Ensemble	Kaks bulgaarlast peeti Kreekas ebaseadusliku telkimise eest kinni.
UT NLP	Kaks bulgaarlast peeti Kreekas ebaseaduslike laagrite eest kinni.

Table 12. The third example sentence on English-Estonian language direction with reference sentence and corresponding translations. The most important differences between translations are highlighted.

The final example in Table 13 shows good and coherent translations for all the models. The word "about" could represent *kohta* as well as *ümbruses* in Estonian, which have different meanings. However, if we rely on the reference translation, then the baseline and UT NLP model translations are not correct and the ensemble model represents the meaning most correctly. It depends on what is the context.

Source	Yes, we have a lot of soldiers about town!
Reference	Sõdureid, jah, liigub meil palju!
Baseline	Jah, meil on linna kohta palju sõdureid!
Baseline + BT	Jah, linnas on palju sõdureid!
Baseline + BT + Ensemble	Jah, meil on linnas palju sõdureid!
UT NLP	Jah, meil on linna kohta palju sõdureid!

Table 13. The fourth example sentence on English-Estonian language direction with reference sentence and corresponding translations. The most important differences between translations are highlighted.

Altogether, translations were of good quality and provided coherent meanings. The differences between translations stood out more, when sentences got longer. When compared to the Estonian-English language pair qualitative analysis, then for English-Estonian direction the translations differed more, however translations remained mostly correct nevertheless. In fact, the example of soldiers, where translations failed on Estonian-English, then *vice versa* the translations from English to Estonian were good. What is more, findings here showed that the better the model was the better quality the translations tended to have as well.

5.3 Future experiments

To raise the quality of Estonian-English language pair translations more, several approaches could be implemented. One of them is for example reranking. Shi et al [10] have showcased in their work that it is a resultative approach for improving the results. In addition, Pinnis et al. [17] have shown that double backtranslation, meaning an additional cycle of backtranslation could raise the quality. Overall, in the future, it would be interesting to train models on other language pairs that contain Estonian and see how they perform.

6 Conclusion

In this work, a new state-of-the-art bilingual machine translation model for the Estonian-English language pair was developed. New benchmark for the Estonian-English translation was set and previous best-performing Tilde's model was outperformed. To achieve the result, the most current technologies and data were used. Additional approaches for results enhancements were backtranslation and ensembling.

At first, two baseline models were trained on the publicly available large amount of parallel data. Additionally, backtranslation was used to produce more data from available monolingual data. New synthetic data was combined with the parallel data and augmented corpus for the baseline models to train on was produced.

Taking the approach of ensembling into account, the more checkpoints were in the ensemble, the better the results for the ensemble got. The best-performing ensembles consisted of six checkpoints and were a combination of baseline and backtranslation models. Results showed that when the ensembles consisted of only the same model's checkpoints then the translation quality was not as high as for the combination of the models' checkpoints. In all cases, ensembling performed better than any of the models alone.

Quantitative analysis of the work showed that both of the approaches that were taken enhanced the BLEU score for both language directions. The differences were not marginal, but positive nevertheless. Although the qualitative analysis was subjective, it still showed in most cases that machine-translated outputs can be of very high quality compared reference translations. On the other side the analysis showed that models are not perfect and can produce incorrect translations as well.

In addition to the practical system developing part of the thesis, a thorough representation of theory implemented in this work was delivered. For future implementations to enhance the quality of translations even more, reranking and double backtranslation were suggested.

References

[1] Papieni, K., Roukos, S., Ward, T., Zhu, W-J. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* Philadelphia: Association for Computational Linguistics, 2002, pp. 311-318. https://dl.acm.org/doi/pdf/10.3115/1073083.1073135

[2] Post, M. A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation (WMT18)*, 2018. arxiv.org/pdf/1804.08771.pdf

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez,A.N., Kaiser, L., Polosukhin, I. Attention is all you need. *In Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010. https://arxiv.org/pdf/1706.03762.pdf

[4] Sennrich, R., Haddow, B., Birch, A. Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin: Association for Computational Linguistics, Vol. 1, 2016, pp. 1715-1725. https://aclanthology.org/P16-1162.pdf

[5] Gage, P. A new algorithm for data compression. *The C Users Journal archive.* Vol. 12, 1994, pp. 23-38. https://www.derczynski.com/papers/archive/BPE_Gage.pdf

[6] Tang, G., Müller, M., Rios, A., Sennrich, R. Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brussels: Association for Computational Linguistics, 2018, pp. 4263–4272. https://arxiv.org/pdf/1808.08946.pdf

[7] Edunov, S., Ott, M., Auli, M., Grangier, D. Understanding Back-Translation at Scale. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, 2018, pp. 489-500. https://arxiv.org/pdf/1808.09381.pdf

[8] Sennrich, B., Haddow, B., Birch, A. Improving Neural Machine Translation Models with Monolingual Data.*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin: Association for Computational Linguistics, Vol. 1, 2016, pp. 86-96. https://arxiv.org/pdf/1511.06709.pdf

[9] Poncelas, A., Shterionov, D., Way, A., Buy Wenniger, G. M., Passban, Peyman. Investigating Backtranslation in Neural Machine Translation. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*. Alacant: European Association for Machine Translation, 2018, pp. 249-258. https://arxiv.org/pdf/1804.06189.pdf

[10] Shi, T., Zhao, S., Li, X., Wang, X., Zhang, Q., Ai, D., Dang, D., Zhengshan, X., Hao, J. OPPO's Machine Translation Systems for WMT20. *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, 2020,

pp. 282-292. https://aclanthology.org/2020.wmt-1.30.pdf

[11] Tiedemann, J. OPUS – Parallel Corpora for Everyone. *Proceedings of the 19th Annual Conference of the EAMT: Projects/Products*. Riga: Baltic Journal of Modern Computing. 2016, pp. 384. https://aclanthology.org/2016.eamt-2.8.pdf

[12] Imamura, K., Fujita, A., Sumita, E. Enhancement of Encoder and Attention Using Target Monolingual Corpora in Neural Machine Translation. *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Melbourne: Association for Computational Linguistics, 2018, pp. 55-63. https://aclanthology.org/W18-2707.pdf

[13] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M. FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneaoplis: Association for Computational Linguistics, 2019, pp. 48-53. https://arxiv.org/pdf/1904.01038.pdf

[14] Kingma, D. P, Ba, J. L. ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION. *3rd International Conference on Learning Representations*. 2015. https://arxiv.org/pdf/1412.6980.pdf

[15] Kudo, T., Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels: Association for Computational Lingusitics, 2018, pp. 66-71.https://arxiv.org/pdf/1808.06226.pdf

[16] Tiedemann, J., Nygaard, L. The OPUS corpus - parallel and free. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon: European Language Resources Association, 2004, pp. 1183-1186. http://www.lrec-conf.org/proceedings/lrec2004/pdf/320.pdf

[17] Pinnis, M., Rikters, M., Krišlauks, R. Tilde's Machine Translation Systems for WMT 2018. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Brussels: Association for Computational Linguistics, 2018, pp. 473-481. https://aclanthology.org/W18-6423.pdf

[18] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: Institute of Electrical and Electronics Engineers, 2016, pp. 2818-2826. https://arxiv.org/pdf/1512.00567.pdf

[19] Dietterich, T. G. Ensemble Methods in Machine Learning. *Proceedings of the First International Workshop on Multiple Classifier Systems*. 2000, pp. 1-15. http://web.engr.oregonstate.edu/ tgd/publications/mcs-ensembles.pdf

[20] Hansen, L. K., Salamon, P. Neural Network Ensembles. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence. Vol. 12, 1990, pp. 993-1001. http://www2.imm.dtu.dk/pubdb/edoc/imm4833.pdf

[21] Bauer, E., Kohavi, R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*. Vol. 36, 1999, pp. 105-139. https://link.springer.com/content/pdf/10.1023/A:1007515423169.pdf

Appendix

I. Dataset

Corpus	Sentence Pairs
bible-uedin	61 981
CCAligned	4 112 295
DGT	5 081 273
ECB	119 511
ELRA-W0154	17 296
ELRA-W0167	47 255
ELRA-W0168	27 897
ELRA-W0215	5325
ELRA-W0218	8321
ELRA-W0265	10 900
ELRC_2682	769 067
ELRC_2922	277
ELRC_2923	394
ELRC_3382	3695
EMEA	1 021 442
EUbookshop	425 993
EUconst	10 086
Europarl	651 236
GNOME	26 009
infopankki	58 410
JRC-Acquis	781 770
KDE4	229 489
MultiCCAligned	4 112 283
OpenSubtitles	12 486 898
ParaCrawl	3 180 464
QED	164 951
Tatoeba	2441
TED2020	23 449
TildeMODEL	2 063 023
Ubuntu	6242
WikiMatrix	243 870
WMT-News	8000
XLEnt	1 761 285
Total	**37 522 828**

Table 14. Parallel training data corpora and according sentence pairs for Estonian-English language pair.

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Kristiina Kuningas**,

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,

   Training the Best Machine Translation Model for the Estonian-English Language Pair,

   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.

Kristiina Kuningas
04/08/2021