University of Tartu
Institute of Computer Science
Computer Science Curriculum

**Raul Erik Kattai**

# Improving Neural Machine Translation Models with Back-translation and Quality Estimation

**Bachelor's Thesis (9 ECTS)**

Supervisor Andre Tättar

Tartu 2021

## Improving Neural Machine Translation Models with Back-translation and Quality Estimation

**Abstract:**

The best machine translation models are on par with human translators as it is becoming increasingly difficult to differentiate between their translations. To produce high-quality results, a translation model requires a lot of training data. However, there exists a limited number of useful bilingual text corpora. By translating a monolingual corpus with a model, a synthetic bilingual corpus can be created. Because of its lower quality, the synthetic corpus can degrade the model and make its output worse. This bachelor's thesis applies a quality estimation model to a synthetic parallel corpus to filter out unsuitable sentence pairs. The resulting dataset is used to fine-tune a machine translation model. The objective is to improve the model with monolingual data.

**Keywords:**

automatic learning, machine translation

**CERCS:** P176 Artificial intelligence

## Neuromasintõlke mudelite täiustamine tagasitõlke ja tõlkekvaliteedi hindamise meetoditega

**Lühikokkuvõte:**

Parimad masintõlke mudelid väljastavad väga kõrge kvaliteediga tõlkeid, mida on raske eristada inimtõlgitud tekstidest. Tõlkemudelist heade tulemuste saamiseks on vaja palju treeningandmeid. Kasulikke kahendkeelseid korpuseid on piiratud koguses ja eriti keelte, mis on vähem levinud, jaoks. Rohkemate treeningandmete saamiseks saab ühekeelsetele korpustele mudeliga juurde genereerida lausete vasteid teises keeles. Selline sünteetiline kahendkeelne andmestik on madalama kvaliteedi tõttu vähem kasulik ja võib mudelit halvemaks õpetada. See bakalaureusetöö rakendab sünteetilisele andmestikule tõlkekvaliteedi hindamise mudelit, mille tulemus aitab paremad tõlked välja valida, et nendega tõlkemudelit edasi treenida. Protsessi eesmärk on kasutada ühekeelseid korpuseid, et masintõlke mudelit paremaks muuta.

**Võtmesõnad:**

tehisõpe, raaltõlge

**CERCS:** P176 Tehisintellekt

# Contents

**1 Introduction** **4**

**2 Technical background** **6**

2.1 Neural machine translation 6

2.2 Transformer model architecture 6

2.3 Word segmentation with byte pair encoding algorithm 7

2.4 BLEU score 7

2.5 Back-translation 8

2.6 Quality estimation 9

**3 Related work** **11**

3.1 Back-translation 11

3.2 Quality estimation models 11

**4 Methodology** **13**

4.1 General approach 13

4.2 Model architecture and development tools 13

4.3 Data preparation and baseline training 14

4.4 Back-translation and baseline fine-tuning 16

**5 Results** **18**

5.1 Quantitative analysis 18

5.2 Qualitative analysis 19

**6 Conclusion** **22**

**References** **23**

# 1 Introduction

Machine learning models that solve complicated problems require a lot of training data. Such is the case with neural machine translation (NMT) models as well. Finding enough high-quality sentence pairs between two languages is often difficult. The complexity increases further when one or both of the languages are not well known in the world.

While there are a few ways to improve a model after it has plateaued during training, like changing its hyper-parameters or architecture, finding more training data is often the easiest. Moreover, if appropriate data does not exist, then it can be created.

For NMT models, the training input is a bilingual corpus - a collection of source and target language sentence pairs. As there exists a lot of monolingual texts, back-translation can be used to produce bilingual corpora. The method involves translating monolingual data with a pre-trained model to create a synthetic bilingual corpus. The result is inferior to a human translated corpus because it contains irrelevant and even destructive data for the model. The unproductive pairs must be filtered out from the synthetic corpus for it to be helpful in training.

Quality estimation (QE) can be a way to filter data. It provides a quality score for the model's output without comparing it to a human-translated sentence. One of the goals of the thesis is to investigate if QE has the potential to filter synthetic bilingual corpora effectively. It also tests the notion that a well-translated synthetic corpus is superior to a poorly translated synthetic one.

The tests were done by creating two filtered datasets from a synthetic corpus using heuristic rules and a QE model's scoring. To identify if the QE filtering process is impactful, an NMT model was trained with both datasets. A baseline model that is only trained on widely accessible bilingual data was created as well. The quality of the final models is indicated and compared in reference to their BLEU score. The chosen language pair is English-German for its abundance in available data.

Chapter 2 explains the relevant technical background. It gives an overview of neural machine translation, Transformer model architecture, byte pair encoding, BLEU score, back-translation, and quality estimation.

Chapter 3 gives an overview of other related works in back-translation and two quality estimation frameworks. It presents a couple of improvements that other papers have achieved with back-translation.

Chapter 4 includes the methodology. It describes what was done with the neural machine translation model and data.

Chapter 5 outlines and compares the quality of the models that were trained on different data. There is an analysis of the experiment and suggestions for future improvements.

# 2 Technical background

## 2.1 Neural machine translation

The field of machine translation utilises software to convert text or speech from one language to another. Before NMT, there were many different approaches, for example, rule-based and statistical machine translation. The translation process in those was word and phrase-based, which meant that a lot of meaning and context was lost. The resulting sentences lacked correct grammar and became more unclear the more prolonged the input was.

NMT translates with neural networks, which have become very popular for being valuable and effective in many tasks. A scientific article by Cho et al. [1] described one of the first successful architectures for sequence transduction. It was an encoder-decoder architecture, where the encoder transforms the input sequence into a fixed-length vector, and from it, the decoder generates an output sequence. In more detail, each symbol read by the encoder creates a hidden state. Its final hidden state is a summary of the whole input. The decoder has separate hidden states, and each one is calculated from its previous hidden state, previously output symbol, and the encoder's final hidden state. The next output symbol is predicted based on the decoder's current hidden state.

The described architecture performed equivalent to a statistical machine translation system while indicating a lot of potential improvement [1]. Because neural networks derive from the whole sentence, more context is kept. They find abstract patterns in data to produce better, more human-like output, although some limitations remain. The models have a finite vocabulary and perform best when restricted to a single domain.

## 2.2 Transformer model architecture

Current state-of-the-art NMT models are based on the Transformer architecture. The Transformer, created by Vaswani et al. [2], utilises self-attention mechanisms while using an encoder-decoder structure similar to recurrent neural networks. Although the attention mechanism has been used before, the Transformer relies solely on attention and does not need recurrence or convolutions.

The authors of the Transformer describe the attention function as a combination of vectors Q, K, and V, where Q is a query and K, V is a set of key-value pairs. Each value V gets a weight by computing the similarity between Q and every K. This way, with a query and multiple key-value pairs, the model can find the most relevant value to the query. Compared to recurrence, where each hidden state depends on the previous hidden state, the Transformer can instead, with attention, query the most relevant states regarding the current state. According to Vaswani et al. [2], self-attention improves three crucial areas: total computational complexity per layer, the computation that can be parallelised, and path length between long-range dependencies in the system. The latter is very impactful in sequence transduction tasks.

Its creators noted that the Transformer could understand syntactic and semantic relations better than its predecessors. Additionally, higher performance and significantly lower training time have made the Transformer architecture widely used.

## 2.3 Word segmentation with byte pair encoding algorithm

Byte Pair Encoding (BPE) algorithm is a data compression algorithm [3]. It finds the most used byte pair in data and replaces it with a not yet present byte. The process is repeated until there are no more byte pairs to replace or there are no unused bytes left to represent a pair.

A neural network's vocabulary size is always limited by the amount of training data and the vocabulary size's impact on the model's performance and quality [4]. At a certain size, a growing vocabulary starts to decrease translation quality. Sennrich et al. [5] proposed a solution that allows the models to perform open-vocabulary translations. To merge recurring characters and character sequences, they applied BPE's fundamentals to words. Their algorithm transforms words into smaller subword units that become the model's vocabulary. Consequently, previously unseen words get separated into subword units that the model can understand. Furthermore, the translation of segmented rare words gets a better accuracy [5].

## 2.4 BLEU score

Evaluating an NMT model's translation quality is an essential part of its development. Evaluation done by humans is a challenging and time-consuming task. Bilingual evaluation understudy

(BLEU) by K. Papineni et al. [6] is the most popular automated quality evaluation metric out of many. The authors state that it correlates highly to the human judgement of translation quality. BLEU metric compares words and their order from the model's output with one or many synonymous reference translations by humans. The score is averaged over multiple sentences to represent an overall similarity to human translation quality and not punish small specific mistakes in sentences.

There are arguments and examples that BLEU is used excessively in the field, and a higher score does not necessarily indicate a better model [7]. Nevertheless, the authors think that BLEU can be irreplaceable in some situations and remains valid for being inexpensive, but there are cases where human evaluation is required instead.

## 2.5 Back-translation

Sennrich et al. [8] have said that the best neural machine translation models are trained with bilingual corpora. Although, they point out that monolingual corpora have been used for several decades in improving solutions to specific translation problems like translating one word or a short phrase. There are many monolingual texts, so finding a practical use for them would be beneficial, as machine learning models usually work better when trained on a larger amount of data. For some requirements, there could not exist enough bilingual corpora. Such data shortage can happen in machine translation of field-specific texts or less common languages.

The previously mentioned article [8] experimented with dummy source sentences and back-translation to use monolingual data. The first method adds monolingual data to an already existing bilingual corpus so that nothing is translated and every added sentence is paired with an empty marking. Their results show that too many of these pairs make the model forget what was previously learned. Forgetting limits the possible usage of available sentences. The second method, back-translation, means translating monolingual data with a trained model to create new parallel corpora [8]. The translated sentence will be the source in the pair, and the original sentence will be the target. In this case, the translation process requires a model that translates from target language to source language. At the same time, the resulting synthetic corpus can be used on a source to target language model. It is worth mentioning that there exists forward-translation, where the original sentence's and its translation's positions in the final

corpus are vice versa. Their results [8] showed that back-translation is a beneficial method. Training on a combined dataset of synthetic bilingual data and human-translated data improved their English to German model +2.8-3.4 BLEU points.

Yang et al. [9] advocate for back-translation but explain that the resulting language pair is still inferior to a human-translated sentence and could impact the model negatively. They think that the problem lies in not making the model prefer human translations while creating an output. They argue that forward-translation can be just as good as back-translation and pose the question of using monolingual data simultaneously on the source and target side. As a solution, the authors propose a new neural machine translation model that uses human and machine translations differently. The new architecture can use monolingual corpora in both the source and target side and do it in higher quantities without losing translation quality.

In this thesis, back-translation was used to create synthetic data. The modification of the model was avoided to make the solution applicable to every machine translation model. Synthetic corpus was filtered with heuristic rules and a QE model to remove low-quality pairs. Filtering with QE should leave only the finest sentence pairs and thus improve the baseline model the most.

## 2.6 Quality estimation

BLEU evaluation still involves a human-translated reference even though the assessment itself is automated. Quality estimation of a neural machine translation does not need human assistance as it bases its evaluation solely on the source and target sentence. It is fully automated and very low-cost.

There are multiple quality estimators, and each one of them has its way of evaluating quality. One indicator of quality can be the uncertainty of a translation. Uncertainty means that there are many equally probable translations to a sentence [10], but not all of them might seem perfect for a human. According to Ott et al. [10], there are two types of uncertainty:

- Intrinsic uncertainty, which means that for one sentence, there are multiple semantically equivalent translations. In addition to paraphrasing, some words might be optional, or some languages could have more specifications like indicating the gender of a noun.
- Extrinsic uncertainty, which is caused by insufficient training data. For instance, the target is a partial translation of a complete copy of the source sentence.

By knowing the uncertainty of each translation, a threshold can be set to remove all sentences that fall below it. Then the sentences that are left, according to the QE model, do not have many other ways to express their meaning and thus are good quality translations.

While QE has greatly advanced in the last couple of years, a study states that current methods are more guessing than estimating [11]. It points out three problems about datasets used to evaluate QE models:

- Good quality translations are in the majority.
- Datasets contain artefacts. They cause the models to find patterns that are present in the train and test data but not in natural language.
- The artefacts correlate well with human judgement scores.

They conclude that QE models assess sentence complexity and fluency but not adequacy. Lack of generalisation makes the current models untrustworthy and highly situational for translation quality evaluation.

# 3 Related work

## 3.1 Back-translation

Back-translation is a popular quality improvement technique. It is used in many papers submitted to the WMT20 machine translation conference [12]. The conference is held annually and establishes different translation tasks for participants to compete in and thus evolve the machine translation field with new state-of-the-art solutions. In WMT20 [12] the state-of-the-art BLEU scores for news translation tasks were 48.0 and 43.9 for English-German and German-English, respectively.

In OPPO's Machine Translation Systems for WMT20 [13], Shi et al. describe their submission for the News Translation task. One of their boosting techniques among fine-tuning, ensemble and reranking was back-translation. They trained systems for all 22 language pairs suggested in the task. The paper gives a detailed overview of required preprocessing steps and specific heuristic filters, most of which are also implemented in this thesis. Their English to German baseline model was trained on all the provided parallel corpora and reached 42.6 BLEU points. However, its BLEU score increased with neither back-translation nor forward-translation. That was not the case for every other language pair.

Samsung R&D Institute Poland submission to WMT20 News Translation Task [14] gives an overview of their submission for the same task. They made models for six language pairs that did not include English-German. However, high-quality synthetic data from back-translation improved the models in every case. For example, English-Polish improved by +0.2 BLEU, Polish-English +0.7 BLEU and Czech-English +3.6 BLEU.

Synthetic data is commonly filtered by heuristic rules or sentence alignment scores [13, 14]. Filtering with QE models is publicly widely unexplored.

## 3.2 Quality estimation models

Here is an introduction of two QE models, from frameworks Crosslingual Optimized Metric for Evaluation of Translation (COMET) [15] and OpenKiwi [16], and their submission to WMT20 and WMT19, respectively. The first QE model was used in this thesis.

The COMET models are based on XLM-R architecture [15]. The framework supports multiple machine translation evaluation models and three metrics: Human-mediated Translation Edit Rate (HTER), Multidimensional Quality Metric (MQM) and Direct Assessment (DA). While the models do slightly differ, the architecture is an intricate combination of a cross-lingual encoder and a pooling layer. There is one model that is purely quality estimation, meaning it does not need reference translations. It is called *wmt-large-qe-estimator-1719*, and it uses DA as a metric. DA represents a human's opinion of a single sentence's quality on a scale of 0-100, where higher means superior quality. The framework's creators submission to WMT20 Metrics shared task showed that their models are competitive or state-of-art in many different tasks and metrics [17].

The second QE model is from OpenKiwi [16]. Kepler et al. submission to WMT19 Shared Task on Quality Estimation had notably better results than other submissions [18]. Their systems built upon the OpenKiwi framework, and they implemented a new Transformer predictor-estimator model. The best system was an ensemble of multiple different models. Its scoring can be applied to both words and sentences. In the first case, each word gets a tag OK or BAD, and between words, there will be label *gaps* if the context is missing. For sentences, the result is a value, which implicates how many edit operations are required to correct it. For best results, this QE solution combines both word and sentence level scoring.

# 4 Methodology

## 4.1 General approach

The upcoming parts of this chapter describe the approach in more detail, excluding the comparison of models in the next chapter. A summary of the experiment is as follows:

1. Collect and preprocess English-German bilingual data.
2. Train a baseline model and a model for back-translation with the bilingual data.
3. Collect and preprocess German monolingual data.
4. Back-translate monolingual data to create a synthetic English-German bilingual corpus.
5. Generate direct assessment scores for each sentence pair in the synthetic corpus with a QE model.
6. Filter the synthetic corpus according to its QE scores and heuristic rules. In total, three differently filtered corpora are made.
7. Fine-tune the baseline model with each filtered synthetic corpus.
8. Compare the baseline model and the three models trained with synthetic parallel data.

Model training, back-translating and QE scoring were done in High Performance Computing Center of the University of Tartu [19]. These processes require a GPU and have to run uninterruptedly for multiple hours or days. A CPU is capable of data preparation, filtering and binarisation, so these were done locally.

## 4.2 Model architecture and development tools

The model used in this thesis is based on the Transformer architecture [2]. Specifically, the model's architecture is Fairseq's [20] *transformer_wmt_en_de_big*. Fairseq is a sequence modelling toolkit written in Python. It contains many useful tools that allow researchers to focus on natural language processing experiments without creating fundamental technical solutions. It was used in this thesis to binarise data in the pre-training phase, train the model using a pre-defined architecture and generate translations with the trained model.

COMET framework's [15] model *wmt-large-qe-estimator-1719* was used in this thesis for quality estimation. There are very few frameworks with QE support, and most of them are not up to

date. Because of it, publicly available pre-trained QE models are rare. To get a suitable QE model for a task, it will most likely have to be trained oneself. That requires thorough data labelling and preprocessing. QE models from COMET [15] and OpenKiwi [16] were tried in this thesis. However, OpenKiwi was incompatible with the training environments, most likely for being slightly out of date, and thus was unusable for scoring.

## 4.3 Data preparation and baseline training

The baseline model was trained on English-German bilingual data. Raw data consisted of 36.5 million sentences. Sentences were taken from ParaCrawl v5.1[1], EuroParl v10[2] and News Commentary v15[3] datasets. ParaCrawl, which makes up most of the dataset, was chosen because it has a mediocre corpus quality. That makes the filtering process more impactful and shows better the importance of data preparation.

Before training, the bilingual data had to be preprocessed and filtered. These two steps were implemented in Python. In the preprocessing step HTML tags, all non-UTF-8 characters and consecutive whitespaces were removed. Sentence punctuation was normalized and tokenised. After preprocessing each sentence, a pair had to pass the following heuristic filters:

1. Source and target sentences are in their respective languages.
2. Source and target are not the same.
3. The pair does not contain repeating words.
4. Source and target are each less than 200 words.
5. The word ratio between source and target is in the range of 0.4 to 2.5.
6. The ratio of characters per word is in the range of 1.5 to 12.
7. No words are over 25 characters long.

The filters were mostly taken from OPPO's Machine Translation Systems for WMT20 [13]. They remove sentence pairs with an irregular structure because those have a higher possibility of being poor translations. For example, repeating words are rare in a correct sentence. Detailed removal statistics are presented in Table 1. The filters were necessary because the original

[1] https://www.paracrawl.eu/index.php
[2] https://www.statmt.org/europarl/
[3] https://opus.nlpl.eu/News-Commentary.php

datasets contain low-quality pairs that might not match at all. In total, 2.2 million sentences were removed, leaving 34.3 million left to train with (Table 2).

Table 1. Number of sentence pairs removed by heuristic filters

| Filter | # removed | # remaining | Retention rate |
|---|---|---|---|
| Initial sentences | 0 | 36 571 052 | 100.00% |
| Long sentence | 159 | 36 570 893 | 100.00% |
| Bad character to word ratio | 1083 | 36 569 810 | 100.00% |
| Source equals target | 3413 | 36 566 397 | 99.99% |
| Not a pair | 12 183 | 36 554 214 | 99.97% |
| Bad word count ratio | 14 255 | 36 539 959 | 99.96% |
| Incorrect language | 63 458 | 36 476 501 | 99.83% |
| Long word | 537 810 | 35 938 691 | 98.53% |
| Repeating tokens | 1 591 932 | 34 346 759 | 95.57% |
| **Total** | 2 224 293 | 34 346 759 | 93.92% |

Table 2. Bilingual training sets in sentence pairs

| Corpus | # total | # removed | # remaining |
|---|---|---|---|
| ParaCrawl | 34 371 306 | 2 057 737 | 32 313 569 |
| Europarl | 1 828 521 | 129 766 | 1 698 755 |
| News Commentary | 371 225 | 36 790 | 334 435 |
| **Total** | 36 571 052 | 2 224 293 | 34 346 759 |

The next step was to train a SentencePiece[4] model with joint vocabulary and apply it to training and validation data. This generated subword units with byte-pair-encoding. The vocabulary size for SentencePiece model training was set to 32K, with 1.0 as character coverage.

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

The model was trained on binarised data created from encoded subword unit sentences. Training parameters were mainly taken from Edunov et al. [21], specifically from the example[5] of their experiment implemented with Fairseq. In the article, they show that sampling is the best way to generate synthetic source sentences. In the thesis, beam search was used instead. Beam search has lower potential but is still able to generate good sentences [21]. Training duration for the baseline English-German model was set to 12 epochs with 0.001 as a learning rate. It took about six days on a single NVIDIA Tesla V100 GPU. A German-English model was trained on the same data to be used later for back-translation. Its training duration was limited to 8 days with the same parameters. During that time, it managed to reach the 16th epoch.

The baseline English-German model scored 30.1 BLEU. That is significantly lower than state-of-the-art models but leaves much potential for back-translation to improve it. The low score is partly due to a small training corpus and its mediocre quality. For English-German, there are hundreds of millions of additional sentence pairs available. However, using that much data to reach state-of-the-art takes months of training and is not feasible in the timeframe of this thesis. Additionally, there is a slight domain mismatch between parallel and back-translated data.

## 4.4 Back-translation and baseline fine-tuning

The German-English model got a BLEU score of 37.9 on the WMT20 test set. The score is not state-of-art but adequate for back-translation. Such a model would produce translations of varying quality, which is an excellent way to test the importance of filtering and see how well the QE model filters. For monolingual corpus, News Crawl 2020, which contains 50 million German sentences, was used. From it, 8.4 million sentences were randomly selected. They were punctuation normalised, tokenised, encoded to subword units with the existing SentencePiece model and binarised. Their translation with a previously trained German-English model generated 8.4 million English sentences. They were put together with the original 8.4 million German sentences to give as input for the filtering processes.

The first synthetic corpus was not filtered. It contained all 8.4 million rows, including the very low-quality pairs. For the second corpus, the heuristic filters described in chapter 4.3 were applied. This filtering process resulted in 8.1 million sentences. The third corpus was based on

[5] https://github.com/pytorch/fairseq/blob/master/examples/backtranslation/README.md

the second corpus because heuristic filters are quick and straightforward and remove many very low-quality pairs. They should be used together with other filtering methods to reach the highest quality dataset. So QE model *wmt-large-qe-estimator-1719* from COMET was applied. It put out a direct assessment score between 0 and 1 for each pair, and if it was under 0.25, then the pair was removed. It was chosen because it kept approximately half of the sentences while being as high as possible. This method removed 4.2 million sentence pairs. The filtering of these three training sets is also depicted in Table 3.

Table 3. Synthetic corpus filtering results in sentence pairs

| **Synthetic dataset** | **# removed** | **# remaining** |
|---|---|---|
| No filters | 0 | 8 415 268 |
| Heuristic filters | 288 463 | 8 126 805 |
| Heuristic filters + QE | 4 249 258 | 3 877 547 |

The generated English side of all three synthetic corpora was encoded with SentencePiece. The German sentences were already processed, for they were used in back-translation. The synthetic corpora were binarised. Then the baseline model was fine-tuned with each of them. The parameters were the same as with the baseline model. The training duration was set to 10 epochs. The outcome was three English-German NMT models trained on different subsets on synthetic bilingual data. Two additional models were trained on a combination of original parallel data with heuristic and QE filtered back-translated data, respectively. Their results show if fine-tuning with back-translated corpus is enough or the synthetic corpus has to be combined with an authentic corpus.

# 5 Results

## 5.1 Quantitative analysis

The fine-tuning did not improve scores. As the baseline model got 30.1 BLEU, fine-tuning with different synthetic corpora ranged from 22.8 to 24.1 BLEU. Among those, the corpus filtered with QE was 23.2 BLEU. However, back-translated sentences combined with the original corpus did reach close to the baseline model score, but did not surpass it. All of the training results can be seen in Table 4.

Table 4. Final BLEU scores of the models

| English-German model training method | Sentence pairs in millions | BLEU |
|---|---|---|
| Training on the bilingual corpus (baseline) | 34.3 | 30.1 |
| Fine-tuning baseline model with raw back-translated data | 8.4 | 22.8 |
| Fine-tuning baseline model with heuristic filtered synthetic data | 8.1 | 24.1 |
| Fine-tuning baseline with heuristic and QE filtered synthetic data | 3.9 | 23.2 |
| Baseline corpus combined with synthetic heuristic filtered data | 34.3+8.1 | 29.2 |
| Baseline corpus combined with heuristic and QE filtered synthetic data | 34.3+3.9 | 29.8 |

These results show that fine-tuning with only back-translated data is not effective and degrades the model. Synthetic data is best used by combining it with a bilingual corpus. Even then, it might not improve the model. An essential match has to be the domain of the sentences. Although the train and test set used in this thesis was from NewsCrawl 2020, the validation set was from NewsCrawl 2019. Within a year, news topics change a lot and the difference can cause the model to sway into an unrelated domain.

Additionally, many parameters can be adjusted, which could improve the BLEU score. For example, the threshold of DA score for filtering out low-quality sentences. To determine all of the best performing parameters without in-depth testing is very difficult.

The BLEU scores indicate that filtering the synthetic corpus is important. Fine-tuning with the corpus that was not filtered got the worst score of 22.8 BLEU. Using the QE model's DA scores to filter did not show better results than simple heuristic filters in fine-tuning. A reason for QE filtered corpus' lower score in fine-tuning could be fewer sentence pairs. It poses a question of where the balance of data quality and quantity is. The QE filtered set outperformed the heuristic filtered set by 0.6 BLEU when combining synthetic data with baseline parallel corpus. Here the higher quality of QE filtered synthetic corpus became more impactful.

For future work, a state-of-art baseline model in terms of BLEU score can be used to see if QE filtering can push the quality higher or is just an alternative in the data processing. Perhaps it is not even possible to reach current top models with QE aided back-translation. For thorough testing, the models can be trained longer. Also, there is a lot more monolingual data available than was used in this thesis. Finally, it is important to train on a combination of synthetic and authentic sentence pairs while keeping their origin of domain similar. Models perform best when trained in a confined domain.

## 5.2 Qualitative analysis

Here are two examples from the test set and how each model translated it. The tags are source sentence (SRC), reference (REF), baseline model (BL), fine-tuned baseline model with raw back-translated data (FT-NF), fine-tuned baseline model with heuristic filtered synthetic data (FT-HF), fine-tuned baseline with heuristic and QE filtered synthetic data (FT-QEF), baseline corpus combined with synthetic heuristic filtered data (BL+QEF).

- **SRC:** A woman in Maine got 500 letters from United Healthcare within five days
- **REF:** Eine Frau in Maine erhielt innerhalb von fünf Tagen 500 Briefe von United Healthcare
- **BL:** Eine Frau in Maine bekam 500 Briefe von United Healthcare innerhalb von fünf Tagen
- **FT-NF:** Eine Frau in Maine bekam innerhalb von fünf Tagen 500 Briefe von United Healthcare
- **FT-HF:** Eine Frau in Maine bekam innerhalb von fünf Tagen 500 Briefe von United Healthcare

- **FT-QEF:** Eine Frau in Maine bekam innerhalb von fünf Tagen 500 Briefe von United Healthcare
- **BL+QEF:** Eine Frau in Maine bekam binnen fünf Tagen 500 Briefe von United Healthcare

This short and straightforward sentence illustrates how similar all the fine-tuned models are. In this case, their three translations are closer to the reference than the baseline model. The BL translation consists of the exact words, but they are in a different order. BL+QEF is closer to the fine-tuned models in terms of structure but includes the word *binnen*, which is not in other translations.

- **SRC:** MSPs were told the legislation needs urgent reform to protect vulnerable people and children.
- **REF:** Den Abgeordneten wurde gesagt, dass dieses Gesetz dringt überarbeitet werden muss, um gefährdete Menschen und Kinder zu schützen.
- **BL:** MSPs wurde mitgeteilt, dass die Gesetzgebung dringend reformiert werden muss, um schutzbedürftige Menschen und Kinder zu schützen.
- **FT-NF:** MSPs wurde gesagt, die Gesetzgebung brauche dringende Reformen, um gefährdete Menschen und Kinder zu schützen.
- **FT-HF:** MSPs wurde erklärt, die Gesetzgebung brauche eine dringende Reform, um gefährdete Menschen und Kinder zu schützen.
- **FT-QEF:** MSPs wurde gesagt, die Gesetzgebung brauche eine dringende Reform, um gefährdete Menschen und Kinder zu schützen.
- **BL+QEF:** MSPs wurde gesagt, dass die Gesetzgebung dringend reformiert werden müsse, um schutzbedürftige Menschen und Kinder zu schützen.

The second sentence shows more variety between the translations. The fine-tuned models are again similar to each other. BL+QEF has the word *gesagt* instead of *mitgeteilt* as in BL. The words appearance shows the impact of the synthetic dataset as *gesagt* is also in two other fine-tuned models' outputs. Apart from it, BL+QEF is closer to BL than the fine-tuned models.

When browsing through all the translations, it is difficult to determine a common mistake each model makes. There is an example of a wrong word, missing word, wrong word order or some

other mistake for each one of them. Because of it, quantitative analysis with BLEU scores is a better indicator of each model's overall translation quality.

# 6 Conclusion

Back-translation is often utilised to create a synthetic parallel dataset from monolingual data to improve NMT models. To reach the best results, filtering out low-quality sentence pairs from the synthetic corpus is essential. This thesis used quality estimation models to decide which sentences to filter out. Such a process is broadly unexplored, so it is interesting to see if QE models are effective filters.

The idea was tested by creating a baseline model with English-German bilingual data. After that, 8.4 million German sentences were back-translated to English to produce a synthetic corpus. Different filters, including a QE model's scoring, were applied to the corpus, resulting in three subsets. The baseline model was further trained to see which one would improve it the most.

The results showed that fine-tuning is not an effective use of synthetic corpus. Fine-tuning lowered the baseline model's BLEU score of 30.1 to around 23. The back-translated corpus should be combined with a bilingual corpus instead. Combined datasets achieved near baseline score but did not surpass it. Filtering with QE did not show distinct advantages over simple heuristic filters. BLEU scores of filtered datasets were higher than not filtered one's. It affirms that a high-quality synthetic corpus is more effective than a low-quality one.

Despite the lack of improvement, back-translation remains a helpful method. Filtering with QE models could still be viable as there are many setups left to explore. The experiment can be improved upon by trying out multiple QE models, using more domain-specific synthetic data and training the models longer.

# References

[1] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 2014. arXiv: 1406.1078 [cs.CL].

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. 2017. arXiv: 1706.03762 [cs.CL].

[3] Philip Cage. A New Algorithm for Data Compression. *The C Users Journal*, 1994, *12*(2), 23-38. ISSN:0898-9788.

[4] Thamme Gowda, Jonathan May. Finding the Optimal Vocabulary Size for Neural Machine Translation. 2020. arXiv: 2004.02334 [cs.CL].

[5] Rico Sennrich, Barry Haddow, Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. 2015. arXiv: 1508.07909 [cs.CL].

[6] Kishore Papineni, Salim Roukos, Todd Ward, Wei-jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, 311-318.

https://www.aclweb.org/anthology/P02-1040/ (07.05.2021)

[7] Chris Callison-Burch, Miles Osborne, Philipp Koehn. Re-evaluating the Role of Bleu in Machine Translation Research. *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

https://www.aclweb.org/anthology/E06-1032/ (07.05.2021)

[8] Rico Sennrich, Barry Haddow, Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. 2015. arXiv: 1511.06709 [cs.CL].

[9] Zhen Yang, Wei Chen, Feng Wang, Bo Xu. Effectively training neural machine translation models with monolingual data. *Neurocomputing*, 2019, *333*, 240-247. ISSN:0925-2312.

[10] Myle Ott, Michael Auli, David Grangier, Marc'Aurelio Ranzato. Analyzing Uncertainty in Neural Machine Translation. 2018. arXiv: 1803.00047 [cs.CL].

[11] Shuo Sun, Francisco Guzmán, Lucia Specia. Are we Estimating or Guesstimating Translation Quality?. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, 6262-6267.

https://www.aclweb.org/anthology/2020.acl-main.558/ (07.05.2021)

[12] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, Marcos Zampieri. Findings of the 2020 Conference on Machine Translation (WMT20). *Proceedings of the Fifth Conference on Machine Translation,* 2020, 1-55.

https://www.aclweb.org/anthology/2020.wmt-1.1/ (07.05.2021)

[13] Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, Jie Hao. OPPO's Machine Translation Systems for WMT20. *Proceedings of the Fifth Conference on Machine Translation,* 2020, 282-292.

https://www.aclweb.org/anthology/2020.wmt-1.30/ (07.05.2021)

[14] Mateusz Krubiński, Marcin Chochowski, Bartłomiej Boczek, Mikołaj Koszowski, Adam Dobrowolski, Marcin Szymański, Paweł Przybysz. Samsung R&D Institute Poland submission to WMT20 News Translation Task. *Proceedings of the Fifth Conference on Machine Translation*, 2020, 181-190.

https://www.aclweb.org/anthology/2020.wmt-1.16/ (07.05.2021)

[15] Ricardo Rei, Craig Stewart, Ana C Farinha, Alon Lavie. COMET: A Neural Framework for MT Evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, 2685–2702.

https://www.aclweb.org/anthology/2020.emnlp-main.213/ (07.05.2021)

[16] Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, André F. T. Martins. OpenKiwi: An Open Source Framework for Quality Estimation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019, 117-122.

https://www.aclweb.org/anthology/P19-3020/ (07.05.2021)

[17] Ricardo Rei, Craig Stewart, Catarina Farinha, Alon Lavie. Unbabel's Participation in the WMT20 Metrics Shared Task. 2020. arXiv:2010.15535 [cs.CL].

[18] Fabio Kepler, Jonay Trenous, Marcos Treviso, Miguel Vera, Antonio Gois, M. Amin Farajian, Antonio V. Lopes, Andre F. T. Martins. Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. 2019. arXiv:1907.10352 [cs.CL]

[19] University of Tartu. UT Rocket [Internet]. share.neic.no; Available from: https://share.neic.no/#/marketplace-public-offering/c8107e145e0d41f7a016b72825072287/ (07.05.2021)

[20] Myle Ott and Sergey Edunov and Alexei Baevski and Angela Fan and Sam Gross and Nathan Ng and David Grangier and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. Proceedings of NAACL-HLT 2019: Demonstrations, 2019.

[21] Sergey Edunov, Myle Ott, Michael Auli, David Grangier. Understanding Back-Translation at Scale. 2018. arXiv:1808.09381 [cs.CL].

# I. License

**Non-exclusive licence to reproduce thesis and make thesis public**

I, **Raul Erik Kattai**,

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 Neural Machine Translation Models with Back-translation and Quality Estimation**,

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.

*Raul Erik Kattai*

***07/05/2021***