UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Dmytro Kolesnykov

Continuous learning for multilingual neural machine translation

Master's Thesis (30 ECTS)

Supervisor: Andre Tättar, MSc
Supervisor: Mark Fišel, PhD

Tartu 2020

Continuous learning for multilingual neural machine translation

Abstract:
With the growing amount of text data, there is also a growing demand for automatic translation systems. The majority of big companies are trying to develop their own translation engines to compete in this field. Especially, there is a need for universal multilingual models that ideally are capable of translating between any languages. This work aims to establish a decent multilingual translation system that continues learning from the monolingual inputs of in-domain data. Thus, to improve the multilingual NMT translation system's performance and transfer knowledge to unseen language pairs without any additional models or parallel data sources. We describe our adaptation of back-translation, a practical approach for data-augmentation, to continuous learning. The results are reported for English, Russian and Estonian languages using only publicly available data.

Keywords: natural language processing, neural machine translation, transfer-learning, back-translation

CERCS: P176 Artificial intelligence

Jätkuv õpe mitmekeelses neuromasintõlkes

Lühikokkuvõte:
Koos pidevalt kasvava tekstiandmete hulgaga on järjest olulisemaks saamas automaatsed tõlkesüsteemid. Enamik suuri ettevõtteid proovivad arendada oma tõlkemootoreid, et sellel alal võistelda. Järjest enam on tähtsamaks muutumas mitmekeelsed masintõlke mudelid, mis oskavad tõlkida kõikide keelte vahel. Selle lõputöö eesmärk on saavutada hea kvaliteediga tõlkesüsteem, mis jätkaks pidevat õppimist domeenipõhistel ühekeelsetel andmetel. Jätkuv õpe aitab tõsta mitmekeelse masintõlke süsteemi headust ja teabesiirde abil õppida tundmatuid keelepaare ilma lisamudeleid treenimata ja paralleelandmeid kogumata. Selles töös kirjeldan tagasitõlke kohandamise moodust jätkuva õppe jaoks - kuidas suurendada paralleelsete andmete hulka sünteetiliselt. Lõpetuseks esitan tulemused inglise, vene ja eesti keele jaoks kasutades ainult vabalt kättesaadavaid andmeid.

Võtmesõnad: loomuliku keele töötlus, tehisnärvivõrkudel põhinev masintõlge, siirdeõpe, tagasitõlge

CERCS: P176 Tehisintellekt

Contents

1 Introduction

Machine translation (MT) is the technology used to translate text between human languages automatically. Although fluent MT can serve as a standalone translation system, it may also be revised manually by post-editors. The key advantages of machine translation comparing to purely human translation are costs and speed. Even though the underlying idea is relatively straightforward, there are many difficulties connected with it. While translation is a natural task for humans, there is no strictly defined way of doing it. Due to the human language's ambiguity and flexibility, there are many possible correct translations for any input, making the evaluation extremely tricky. Modern machine translation is quite impressive because it can be developed without adding any rules or task-specific constraints for translating text from one particular language to another, but rather knows how to translate in general. Therefore, such a system's main task is to learn the parameters that convert the sequence of source words into the sequence of target words directly from the text corpus.

Neural machine translation (NMT) based on encoder-decoder architecture (Sutskever et al., 2014; Bahdanau et al., 2015) has been established as a state of the art in machine translation evaluation reports (Barrault et al., 2019, 2020) beyond traditional approaches like statistical machine translation (SMT) and based solely on neural networks. One of the main advantages of NMT compared to previous industry standards is producing comparable or better results without the need to optimize multiple independent models and relations between them. This leads to the simplification of training pipelines and the ability to obtain end-to-end solutions. Still, NMT has some drawbacks (Koehn and Knowles, 2017). For instance, out-of-domain NMT shows much lower results by sacrificing adequacy for the sake of fluency.

In previous years, neural machine translation (NMT) has gained a lot of attention (Wu et al., 2016; Vaswani et al., 2017) among researchers due to rapid change in the deep-learning field that brings promising improvements. Nowadays, the amount of time needed to train a fairly good NMT system using modern NVIDIA GPUs takes around 3 days, depending on the utilized toolkit (Domhan et al., 2020). Traditional approaches (Bahdanau et al., 2015) involve training a separate model for each translation direction and might still be impractical for the production (Arivazhagan et al., 2019) because the number of translation directions grows quadratically. However, once trained, the time needed for NMT model to generate a translation is quite reasonable (Junczys-Dowmunt et al., 2016).

Naturally, these facts push the research field towards the idea of building the multilingual model capable of translating between many languages (Arivazhagan et al., 2019) or transfer knowledge (Tan et al., 2019) from individually trained unidirectional models. Multilingual NMT can be designed to perform one-to-many (Dong et al., 2015), many-to-one (Zoph and Knight, 2016), bi-directional (Niu et al., 2018) or many-to-many (Firat et al., 2016; Luong et al., 2016; Johnson et al., 2017) translations. An intuition behind creating multilingual NMT is that the learning signal from one language should benefit the quality of other languages (Caruana, 1997). Under this assumption, introducing more languages may allow the multilingual system to generalize better, even in previously unseen (zero-shot) directions (Johnson et al., 2017). Nevertheless, with all else unchanged, multilingual models tend to underperform separate models and usually end-up with poor zero-shot translations when many languages are combined. One way of solving the first issue is to enlarge the model capacity (Zhang et al., 2020). The algorithm for improving zero-shot translations at scale has been recently proposed in the same article.

The premise for the effectiveness of NMT is the availability of aligned parallel data, which is practically costly to collect. Since this fact certainly limits the translation system's scalability, many techniques for the extracting or synthetic generating of parallel data were previously introduced. In particular, the back-translation of monolingual data, which is available in much larger amounts, has been proven effective for this purpose. We describe related work in Section 3 and our adaptation of this simple yet effective approach in Section 5.3.

Consequently, this work's main objective is to further back-translation idea and leverage in-domain monolingual data with continuous learning (1) in a multilingual NMT setting. We hypothesize that the back-translation effectiveness can be extended by increasing the number of update cycles while the amount of monolingual data is fixed. Results described in this work show how often one might retrain the existent model to obtain substantial improvements comparing to back translating all available monolingual data at once.

Research questions:

1. Does continuous learning offer improvement over one-time back-translation and which granularity is better?

2. How does continuous learning impact zero-shot translations of multilingual model?

2 Technical background

Translation is a sequence-to-sequence modeling problem and formally equivalent to finding a target sentence $Y = (y_1, y_2, ..., y_m)$, given a source sentence $X = (x_1, x_2, ..., x_n)$, so that conditional probability of Y is maximazed i.e:

$$\arg\max_{y} p(Y|X) \tag{1}$$

The following section formulates how this problem is solved in the context of NMT and describes chosen approach for training the baseline model.

Encoder-Decoder. Conditional probability (1) can be parametrized by θ with encoder-decoder architecture and jointly trained to fit the parallel corpus $D^P = \{(X, Y)\}_{k=1}^{N}$:

$$\theta^{\star} = \arg\max_{\theta} \sum_{(X,Y)\in D^P} \log\ p(Y|X;\theta), \tag{2}$$

where $\theta^{\star}$ is optimal set of model parameters and $p(Y|X;\theta)$ is factorized using the chain rule:

$$p(Y|X;\theta) = \prod_{i=1}^{m} p(y_t|Y_{1:t-1}, X;\theta) \tag{3}$$

After vocabulary V is built from the training data D^P, each token from the source $x_i \in X$ and target sequence $y_i \in Y$ is represented with corresponding one-hot encoded id vector $x_i, y_i \in \{0,1\}^{|V|}$. Plain encoder recurrent neural network (RNN) is aimed to map variable-length input sequence of tokens X into fixed-length vector representations (i.e., embeddings or hidden states) by consistently updating a hidden state of the recurrent unit for each token in the sequence. Current hidden state h^i is computed from the previous hidden state h^{i-1} and the current input x_i.

$$h^i = f(h^{i-1}, x_i), \tag{4}$$

where f is a non-linear activation function. Then, given the embeddings sequence, the encoder summarizes the whole sentence, for instance, with the last hidden state vector. Then, the decoder outputs one token at a time, conditioning on the input vector and previously generated tokens.

This simple approach already yields good results but has some known limitations with long sequences that are solved to a certain degree with attention mechanisms (Bahdanau et al., 2015). Later, it was further improved with Transformer (Vaswani et al., 2017) architecture that does not use RNN and solely based on attention layers in the encoder and decoder. The Transformer is taken as the main architecture for this work. We use the implementation from Sockeye 2 (Domhan et al., 2020), the basis toolkit for our experiments.

Tokenization. The process of grouping the sequence of characters from the text into some semantically meaningful units (i.e., tokens) is considered an essential step for every MT pipeline. The most straightforward approach to sequence-to-sequence modeling with NMT is dividing the input text into a sequence of word-level units. While practically the amount of different words is infinite, translation system vocabulary is limited. As mentioned by Luong et al. (2015) to the softmax's computationally intensive nature, NMT systems often limit vocabularies to be the top 30K-80K most frequent words in each language. The problem with word-level translation is the necessity to generate a special <unk> token for the unseen words during processing. This introduces the problem of unseen words that will retain an <unk> token regardless of its meaning and further complicates the translation capabilities since every unknown word will be internally represented with a single token that cannot express word uniqueness. On the other hand, using purely character-level segmentation is suboptimal for alignment with the attention mechanism. One common strategy to tackle mentioned issues is to apply segmentation with subword units (Sennrich et al., 2016a), assuming that rare words could actually be translated within smaller parts.

Byte Pair Encoding. BPE (Gage, 1994) is a data compression technique that can be applied for subword segmentation. These segments can be extracted automatically from the corpus. The segmentation algorithm starts with initializing character vocabulary. Then, iteratively merge the most frequent pair of neighboring characters "a", "b" and replace them with a new symbol "ab" until a fixed number of merge operations is completed. Produced symbols represent the most frequent character n-gram, and their amount plus the initial character number forms the size of the vocabulary. This way, segmentation achieves a trade-off between vocabulary size and the number of symbols required to encode the text (length of token sequences is minimized). Also, this resolves the problem of rare words since during inference algorithm applies learned merge operations on separated word characters. Thus, common words will be represented as one symbol, whereas words with rare character combinations will be divided into smaller subword units or characters.

Back-translation. The main idea of back-translation is to utilize monolingual data without changing the model's architecture. It is accomplished by automatically translating the monolingual target data to the source language using the target to source model. These translations are then used jointly with the original target text to form additional bitext data for the primary model. Such data is called synthetic, and back-translation can be considered as a data augmentation technique.

While back-translation is usually used as a heuristic within the lack of parallel data, it can be derived from a statistical perspective. Target-side monolingual data can estimate the prior of the target sentences. The NMT model's optimization requires the empirical joint distribution of source and target sentence pairs obtained from the bilingual corpus. One way of using it is to train a separate language model and integrate it with the existing translation model (Gülçehre et al., 2015). However, with the Bayes rule, the desired conditional probability can be decomposed into the language probability (prior) and reverse translation probability. In the context of NMT, where the decoder can already condition on the target side text, the only component that is missing is the reverse translation probability. This probability can be approximated with the empirical distribution of synthetic data obtained from back-translation. This approximation quality will depend on the target-to-source model adequacy and the generation algorithm's choice (e.g., beam search or sampling). This way, monolingual data can be leveraged without changing NMT architecture.

Evaluation. To test the models, we report BLEU (Papineni et al., 2002) calculated with the sacreBLEU (Post, 2018) implementation, a metric for automatic evaluation that measures overlap between translations and references. For this, every input is preprocessed, translated, and then detokenized for the assessment. According to Bogoychev and Sennrich (2019), authors of the original back-translation, BLEU is very sensitive to the choice of data augmentation. Models trained with back-translation excel when the input to the translation system is itself a human translation, and the original text is used as a reference. The gain on the artificial half of the test set can be big enough to prevail in the aggregated results. Thus, we separate artificially reversed references during testing of models fine-tuned on back-translation to capture the translationese effect.

3 Related Work

The first successful attempt to demonstrate back-translation effectiveness in the case of NMT (Sennrich et al., 2016b) has shown significant improvements in the translation model's quality and adapted it to the new domain. These results were obtained by mixing synthetic data with original human-translated parallel text without distinguishing between them. Based on these findings, many works have been done to further these results and extend back-translation usage.

In particular, there are two generation procedures widely used in recent works: beam-search and sampling. While sampling (Edunov et al., 2018) and noising (Wu et al., 2019) claimed to produce a richer training signal than deterministic beam-search, another possible reason could be that noise makes the model classify synthetic data and able to separate helpful and harmful signal (Caswell et al., 2019) from the training data.

There are a few more known ways of exploring the usage of back-translated data. For example, it can be used as a standalone data-set or in a combination of parallel data with different proportions. While at first glance, the idea to build an NMT system with good performance using only pseudo parallel data seems unfeasible, some works show the opposite results (Park et al., 2017; Poncelas et al., 2018). On the other hand, the hybrid model that uses both actual and artificial data, back-translation can be useful only up to some extent. Since pseudo parallel data quality is usually worse than real human-translated data, monolingual data can also degrade the model performance. The work by Poncelas et al. (2018) investigates this phenomenon and shows the optimal synthetic-to-authentic ratio (2:1), which we will use for our experiment.

BT has been proved to be more or less effective in all (low-resource, mid-resource, high-resource) scenarios. However, for each of them, there are different nuances. For example, when applying BT to the strong baseline, the model can unlearn useful parameters if the synthetic-to-authentic ratio is too high. Some recent work shows this issue can be tackled by explicitly pointing out the model when data is synthetic by adding a unique tag to the back-translated source sentences (Caswell et al., 2019; Marie et al., 2020). On the other hand, in a low-resource setting, when only low-accuracy machine translation systems can be used for the generation, pseudo-parallel data can be filtered (Imankulova et al., 2017) to boost the performance of the source-to-target model.

Furthermore, the NMT model's iterative training with back-translation was previously described by Hoang et al. (2018) and proved to be useful in low-resource and high-resource settings. The main idea of iterative back-translation is as follows: if back-translation helps to obtain a better model, then one might use that same system to produce even better translations for the next step of back-translation and repeat this process until convergence or other stopping criteria. While the method described above is a complement to ours, substantial differences of this work persist in a few important aspects: (i) another language set (ii) the absence of an auxiliary model for target-to-source translations — only one multilingual model is used to perform BT for every direction (iii) different NMT model architecture (iv) several monolingual data partitions of different sizes are used to discover the optimal number of iterations (v) instead of training from scratch we continue training of the baseline.

Other types of semi-supervised approaches also exist for NMT. Dual learning (He et al., 2016) represents the task of training a bi-directional translation model as a two-agent communication game that is solved through the reinforcement learning process. Self-learning with forward-translation (Bogoychev and Sennrich, 2019) is also used, but it is more sensitive to the quality of the system used to produce synthetic data.

Nowadays, back-translation has already become an essential part of modern NMT. Even though it is still an open challenge because there are many unknown factors regarding the effects it introduces to the NMT system.

4 Data

WMT is a workshop that organizes a collection of shared tasks related to machine translation, where researchers compare their techniques against those of others in the field using a common test set. All training data used in this work was provided by WMT for the news translation shared task. This section aims to cover the data used for each stage of the experiments as well as technical details connected with processing it.

4.1 Sources

Three languages were chosen for further investigation: EN(English), RU(Russian), ET(Estonian). The English-centric datasets for training baselines are described in Table 1.

Europarl corpus (Koehn, 2005) Release v7 is extracted from the European Parliament's proceedings from 1996 to 2011 and includes versions in 21 European languages, which we used for training EN↔ET baselines. In order to train EN↔RU baselines, we used The United Nations Parallel Corpus v1.0 (Ziemski et al., 2016). It is composed of human translations of official records and other parliamentary documents of the United Nations (1990 to 2014). Translations are available for six official languages: Arabic, Chinese, English, French, Russian, and Spanish. Paracrawl corpus collected by Bañón et al. (2020) mainly focuses on all 24 official EU languages (including Irish, Maltese, and Croatian) but also targeted Russian and some other languages. It was mined from the collection of web pages in HTML and files in PDF format, using text where available and optical character recognition otherwise. We use Paracrawl as a data source for both EN↔ET, EN↔RU baselines. We clean the training data so that if any of the parallel sentences is empty, contains more than a hundred tokens, or one of the sides has nine times more tokens, then the pair is removed.

Table 1. Baselines training data

language(s)	dataset(s)	samples
EN↔ET	European Parliament Proceedings v7	$\approx 0.65m$
	ParaCrawl v7.0	$\approx 2.85m$
	Total	3.5m
	Filtered	**3m**
EN↔RU	The United Nations v1.0	$\approx 23.25m$
	ParaCrawl v7.0	$\approx 5.38m$
	Total	28.6m
	Filtered	**26.8m**

For the experiments with back-translation, we employ monolingual news data referred to in Table 2. We filter monolingual data by pre-trained fastText [1] language detection model (Joulin et al., 2016a,b). Then, sixteen million lines per language are randomly sampled and accumulated into ninety-six million synthetic parallel data lines by translating selected monolingual data for each language into every other language. In our case, we have chosen three languages that lead to six possible translation directions. We choose the amount of data so that all synthetic data can be seen during approximately one day of training.

Evaluation and testing sets for EN↔ET are both taken from WMT18 (Bojar et al., 2018). For EN↔RU the evaluation set is taken from WMT19 (Barrault et al., 2019), and tested on WMT20 (Barrault et al., 2020). For testing the performance of RU↔ET

[1] `https://github.com/facebookresearch/fastText/`

Table 2. Monolingual data

language(s)	dataset(s)	samples
EN	News Commentary v15	$\approx 0.6m$
	News Crawl 2007-2019	$\approx 233.5m$
RU	News Commentary v15	$\approx 0.4m$
	News Crawl 2008-2019	$\approx 93.8m$
ET	BigEst	$\approx 40.4m$
	News Crawl 2014-2019	$\approx 5.3m$

zero-shot translations, we use ACCURAT balanced test corpus (Skadins et al., 2010; Rikters et al., 2018).

Table 3. Evaluation data

language(s)	dataset(s)	samples
EN→ET	WMT18/dev	2000
ET→EN	WMT18/dev	2000
EN→RU	WMT19/test	1997
RU→EN	WMT19/test	2000

Table 4. Test data

language(s)	dataset(s)	samples
EN→ET	WMT18/test	2000
ET→EN	WMT18/test	2000
EN→RU	WMT20/test	2002
RU→EN	WMT20/test	991
ET→RU	ACCURAT	512
RU→ET	ACCURAT	512

4.2 Pre-processing

Before introducing raw text to the translation system either for training or evaluation, we perform some preliminary processing steps described in this subsection.

Table 5. True-casing examples

source	Noah was rushed by ambulance to a local hospital.
true-cased	Noah was rushed by ambulance to a local hospital.
source	Four members of the Kemerovo group arrested in Estonia and Spain.
true-cased	four members of the Kemerovo group arrested in Estonia and Spain.

4.2.1 True-casing

True-casing [2] is one of the pre-processing steps that solves the ambiguity of the word casing. It is aimed to convert the capital letter of common nouns at the beginning of the sentences into lower case. On the other hand, proper nouns that should always be written from the capital letter should remain unchanged. In order to decide which words at the beginning of the sentence should remain intact, the frequencies of the words in the whole corpus are calculated. If the word has been written more frequently from the capital letter or has never been encountered before, it is supposed to be left unchanged. True-casing is applied to every data set: parallel, monolingual, evaluation, and test sets. This way, the translation system receives already "true" word casings, regardless of the position in the sentence. As a result, we avoid encoding the common nouns into two different representations, one starting from the capital letter and another from the lower case letter.

4.2.2 Subword segmentation

We employ a similar method to BPE segmentation (Sennrich et al., 2016a) implemented in SentencePiece [3] that shares the same idea but can augment training data with on-the-fly subword sampling from multiple segmentations and their probabilities using a unigram language model (Kudo, 2018) in contrary to deterministic BPE. Segmentation with the unigram language model results in a combination of words, subwords, and character segmentation. The framework treats whitespace as a regular character and introduces a special underscore symbol (U+2581) to solve detokenization ambiguities.

SentencePiece model is jointly trained with vocabulary size 32K and character coverage 0.9995. Obtained vocabulary is passed directly to the translation system, and samples that contained out-of-vocabulary tokens after segmentation were removed before training. Originally there were 3778 distinct symbols before filtering of the whole corpus and 195 afterward. A couple of segmentations with out-of-vocabulary symbols are highlighted in Table 6.

[2] https://github.com/TartuNLP/truecaser

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

Table 6. Subword segmentation examples

source	relaxation in a bath house at the lake Bruņķītis.
tokenized	_relax ation _in _a _bath _house _at _the _lake _B ru **ņķī** tis .
source	...reprinted by the Nestlé Foundation.
tokenized	..._re print ed _by _the _N est l **é** _Foundation .

5 Experiments

In the following chapter, we outline specifics of the setup for training the baseline models and cover the multilingual model fine-tuning method based on back-translation.

5.1 Model hyperparameters

For all results to be comparable, the same default Sockeye architecture (Hieber et al., 2017, p. 13) is employed. Specifically, the base Transformer with six layers of 512 hidden units and eight attention heads for both encoder and decoder. There are also 2048 hidden units for feed-forward layers. Source factors embedding size is set to 8. Each transformer building block is pre-processed with layer normalization, and post-processed with a dropout equals to 0.1 followed by residual connections operation. Translations for evaluation are generated using beam-search of size 5. Back-translations are generated with beam size equals 2.

5.2 Baselines training

Given parallel data, a separate uni-directional (EN $\rightarrow$ RU; RU $\rightarrow$ EN; EN $\rightarrow$ ET; ET $\rightarrow$ EN) as well as bi-directional (EN $\leftrightarrow$ RU; EN $\leftrightarrow$ RU) models are trained for each possible translation direction to compare the performance with the main many-to-many (EN $\leftrightarrow$ RU $\leftrightarrow$ ET) multilingual model which is later picked for back-translation. One way to train a multilingual NMT without changing the model architecture (Johnson et al., 2017) is to add an artificial tag at the beginning of the input sentence to bind translations into the required target language. We used a similar approach but with adding a language tag to each token from the source sentence as a source factor (Sennrich and Haddow, 2016). Thus, for training bi-directional and multi-way translation systems, available bilingual data is reversed and concatenated while the translation direction at both training and evaluation time is specified as an additional feature (Tättar et al., 2019). Every baseline gets a shared vocabulary of subwords from the trained SentencePiece model described in Section 4.2.2.

Two NVIDIA Tesla V100 GPUs were used for training with batch size set to 12K tokens (maximum possible value is 6K per GPU). Model checkpoints are saved every

2,000 updates, and early stopping is triggered after 18 checkpoints without improvement on the validation set. Beyond that, the multilingual model was limited to five days of training, while smaller models with up to 3 days. The learning rate scheduler is plateau-reduce which keeps initialized value 0.0002 until validation metric has not been improved for eight checkpoints. Then, the learning rate is reduced by multiplying on reduce factor 0.8 and restores model weights from the best checkpoint.

5.3 Fine-tuning

Algorithm 1: Continuous learning

Input:

Pre-trained multilingual model, Θ
Target language set, L
Number of back-translation steps, N
Monolingual data, $D^M = \bigcup_{l \in L} D^m_l$

$i \leftarrow 0$;
while $i < N$ **do**
  $D^p_l \leftarrow \emptyset$;
  for $\forall l \in L$ **do**
    $B^l_{size} \leftarrow \frac{|D^m_l|}{N}$;
    Sample B^l from D^m_l ; // $n(B) = B^l_{size}$
    $D^m_l \leftarrow D^m_l \setminus B^l$;
    for $\forall(l' \in L) \wedge (l' \neq l)$ **do**
      $D^p_l \leftarrow D^p_l \cup [\Theta_{translate}(B^l, l') = \{(\hat{x}, y, l) : \forall y \in B^l\}]$;
  $D^P \leftarrow \bigcup_{l \in L} D^p_l$;
  $\Theta' \leftarrow \Theta_{learn}(D^P)$;
  $i \leftarrow i + 1$; // Back-translation iteration is over

Output: Updated model Θ'

Continuous learning. Plain back-translation uses a pre-trained target-to-source model to produce translations from the monolingual data and create parallel data, where the source side is formed from the translations and the target side from the corresponding inputs to these translations. Usually, back-translated data is mixed with parallel data and used to train the model from scratch. Compared to these practices, we do not employ any additional models and continue training the pre-trained multilingual baseline as in (Freitag and Al-Onaizan, 2016) but only on the synthetic data and with several intermediate updates. Since back-translation is applied iteratively, continuation reduces the burden of retraining the baseline on authentic data for every new portion of the artificial data.

Experiment details. To experiment with the optimal number of iterations for continuous learning and to keep results comparable, it is crucial to perform exactly one epoch of training for each chunk of monolingual data. The learning rate for each back-translation iteration is adjusted with a value from the previous step. Arguments to the training loop are the same as for baseline except more frequent saving of the model weights for the fine-tuning stage which is set to 500 updates per interval.

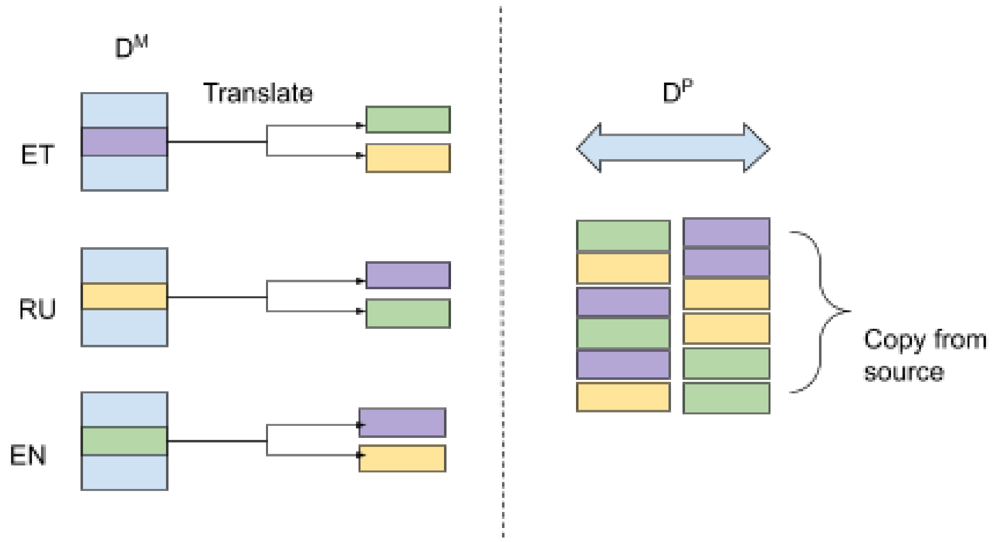

Figure 1. One iteration of back-translation: batch of monolingual data for all languages is translated into every other and then combined into parallel set.

During decoding with a multilingual system, some of the inputs are translated into the wrong language ignoring provided source factors. This especially stands out for zero-shot translation directions. Since the encoder is shared across all languages, the off-target problem is ignored while generating synthetic data.

As described in Section 4.1, the amount of monolingual data used for back-translation is 16m of sentences per language. If the number of back-translation cycles equals 1, then every sentence for each language is back-translated by the multilingual baseline into every other language generating 96 million parallel sentences. If the number of iterations equals 2, then every second sentence is back-translated for the first iteration. The remaining half is back-translated with an already updated model. In such a manner, we applied the continuous learning procedure (1) for a different number of iterations $N = \{1, 2, 4, 8\}$

Table 7. Baseline test results

Test set	Direction	Baseline	BLEU
WMT20	en $\rightarrow$ ru	uni-directional	**18.4**
		bi-directional	17.7
		multilingual	17.5
	ru $\rightarrow$ en	uni-directional	**30.1**
		bi-directional	29.2
		multilingual	29.0
WMT18	en $\rightarrow$ et	uni-directional	17.5
		bi-directional	**18.0**
		multilingual	16.5
	et $\rightarrow$ en	uni-directional	**27.2**
		bi-directional	25.4
		multilingual	24.4
ACCURAT	et $\rightarrow$ ru	multilingual	1.9
	ru $\rightarrow$ et	multilingual	2.2

that overall uses the same amount of data but corresponds to a different batch size per update cycle $B^l_{size} = \{16m, 8m, 4m, 2m\}$. One iteration of continuos learning is schematically illustrated in Figure 1.

6 Results

The BLEU scores for the baseline models are shown in Table 7. For directions with the English target language (which prevails in overall text quantity), the BLEU score is much higher than for other target languages. Secondly, when more languages are accommodated into the model of the same capacity, the performance drops. Thus, our motivation is to improve the multilingual baseline without changing the architecture or retraining it.

As can be seen in Figure 2 and Figure 3, there are completely different BLEU scores when translating original sentences and translationese. Multilingual baseline for English-Estonian (WMT18) case producing a better result with a larger margin (≈ 2 BLEU) on ET$\rightarrow$EN direction given original sentences as a source. As for Eglish-Russian (WMT20) test case, the performance on original test sentences is higher for RU$\rightarrow$EN direction but lower for EN$\rightarrow$RU with a considerable margin (≈ 9 BLEU). Thus, BLEU scores are much higher when original sentences were in Russian either used as a reference or as a source text, while EN$\leftrightarrow$ET translation directions are more stable to this effect.

In both tests on original and translationese sources, dividing data into more batches and reiterating does not show the expected performance boost. On the contrary, when testing on translationese, it is more advantageous to perform only one iteration in terms of performance and complexity. The only case of gaining higher BLEU from back-translation while testing on original translations is the WMT18 English-Estonian evaluation set with the best improvement of $\Delta = 1$ BLEU points, which makes the multilingual model comparable to the unidirectional model for the same translation direction. Otherwise, model performance drops with employing monolingual data.

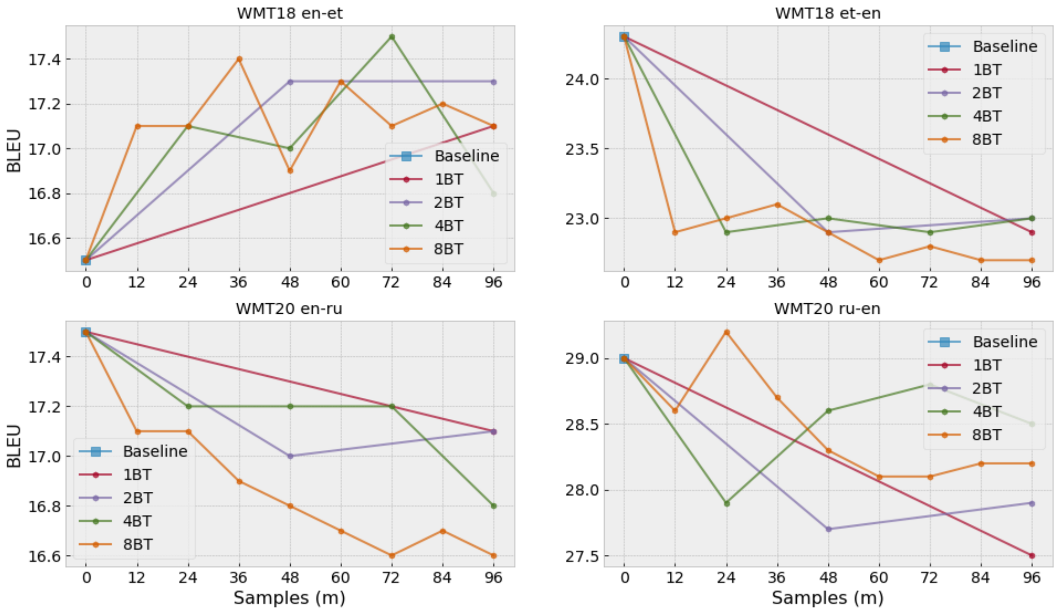

Figure 2. Test on original translations with WMT datasets

However, for zero-shot translations results shown in Figure 4, it is beneficial to re-iterate back-translation with smaller batches of monolingual data. There is no substantial difference in performance between one iteration on all available monolingual data or 1/8 part of it. Curves that represent a higher number of iterations are getting steeper with adding more data. The best results for zero-shot translations are produced with models assigned to a maximum number of back-translation iterations and converging at half of the available samples.

From the translation system output given in Table 8, it can be seen that model with back-translation outputs more complex words endings than a baseline, like *partner*$\rightarrow$ *partneri[le], protsendi* $\rightarrow$ *protsendi[list]*, or *aasta* $\rightarrow$*aasta[ks]* . The Estonian language has many grammatical cases and different endings, which are important for the sentence's general meaning. While baseline is more conservative to put endings, a model based on back-translation adds them more aggressively. Fine-tuned model succeeded at comitative

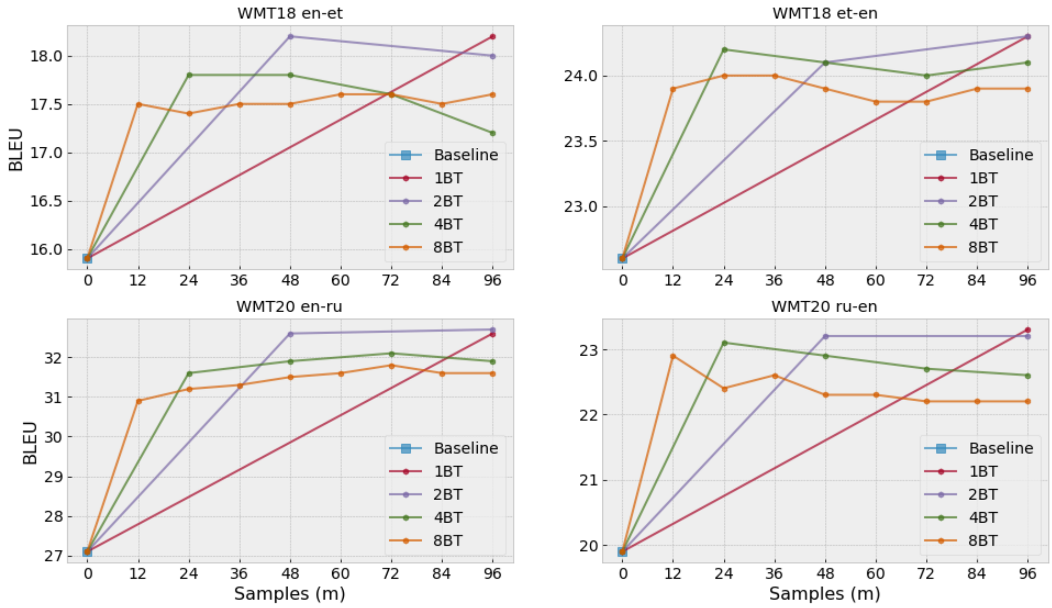

Figure 3. Test on translationese with WMT datasets

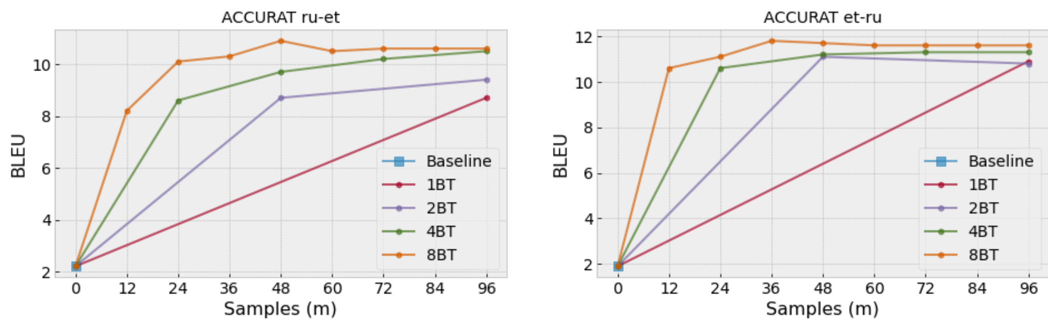

Figure 4. Test zero-shot directions with ACCURAT dataset

case of plural *koer[ad]* $\rightarrow$ *koer[tega]* but makes a mistake to preserve the negation meaning of the phrase: oli muut[mata] $\rightarrow$ muut[usid] $\rightarrow$ ei ole muut[unud]. Both baseline and BT confused the correct pronouns like "kes" (who) instead of "mis" (which) and misinterpreted impersonal verbs.

Interestingly, in Table 9 "USA dollarit" was compressed into one "$" symbol, and the BT model mentioned that "The New York Times" is actually a newspaper, while it was not mentioned anywhere in the source text. Since the baseline was fine-tuned with back-translation on the news domain monolingual data, it presumably learns the context around this entity.

7 Conclusion

In this work, we developed multiple neural machine translation models to explore the application of back-translation in the multilingual, high-resource setting. For this, we train a multilingual baseline able to translate between any direction across English, Russian, and Estonian languages by concatenating all available parallel data. We compare the performance of multilingual baseline with uni/bi-directional baselines to report its initial capabilities. Then, we discover the advantages and limitations of applying continuous back-translation with consequent model updates. We experiment with a different number of update cycles for the fixed amount of monolingual data to achieve this.

Answering research questions:

1. Does continuous learning offer improvement over one-time back-translation and which granularity is better?
2. How does continuous learning impact zero-shot translations of multilingual model?

Q1 Comparing the performance of enhanced models depends on choosing the directionality of the evaluation set. When the input to the model is an original sentence and human translation is used as a reference, in most times, baseline outperforms fine-tuned model. On the other hand, when the input sentence is itself a translation, and the original sentence is used as a reference, every fine-tuned model outperforms the baseline, but more frequent iterative updates are abundant.

Q2 For zero-shot translation directions that were not presented to the baseline directly, continuos back-translation with higher granularity achieves constant improvements. The results show that the best configuration is to divide 16m of monolingual data per language into eight batches and get the gain of 10 BLEU for zero-shot directions.

We conclude that for a strong enough multilingual baseline, the safest strategy to leverage continuous learning is to improve the performance of zero-shot translation directions. For this, the amount of monolingual data can be reduced without loss in performance, and translation into pivot languages can be omitted. Finally, the BLEU metric is ambiguous, and other types of automatic or manual evaluation are essential to fully understand the effects of back-translation on the translation system.

References

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Loïc Barrault, Ondrej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In *WMT (2)*, pages 1–61. Association for Computational Linguistics, 2019.

Loïc Barrault, Magdalena Biesialska, Ondrej 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 Ljubesic, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In *WMT@EMNLP*, pages 1–55. Association for Computational Linguistics, 2020.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *NMT@ACL*, pages 28–39. Association for Computational Linguistics, 2017.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *AMTA*, pages 110–115. Association for Machine Translation in the Americas, 2020.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang

Macherey, Zhifeng Chen, and Yonghui Wu. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019, 2019.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. Is neural machine translation ready for deployment? A case study on 30 translation directions. *CoRR*, abs/1610.01108, 2016.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Multilingual neural machine translation with knowledge distillation. In *ICLR (Poster)*. OpenReview.net, 2019.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *ACL (1)*, pages 1723–1732. The Association for Computer Linguistics, 2015.

Barret Zoph and Kevin Knight. Multi-source neural translation. In *HLT-NAACL*, pages 30–34. The Association for Computational Linguistics, 2016.

Xing Niu, Michael Denkowski, and Marine Carpuat. Bi-directional neural machine translation with synthetic parallel data. In *NMT@ACL*, pages 84–91. Association for Computational Linguistics, 2018.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *HLT-NAACL*, pages 866–875. The Association for Computational Linguistics, 2016.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. In *ICLR (Poster)*, 2016.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351, 2017.

Rich Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, 1997.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *ACL*, pages 1628–1639. Association for Computational Linguistics, 2020.

Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. In *ACL (1)*, pages 11–19. The Association for Computer Linguistics, 2015.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016a. doi: 10.18653/v1/p16-1162. URL https://doi.org/10.18653/v1/p16-1162.

P. Gage. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38, 1994.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535, 2015. URL http://arxiv.org/abs/1503.03535.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040/.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

Nikolay Bogoychev and Rico Sennrich. Domain, translationese and noise in synthetic data for neural machine translation. *CoRR*, abs/1911.03362, 2019. URL http://arxiv.org/abs/1911.03362.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *ACL (1)*. The Association for Computer Linguistics, 2016b.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1045. URL https://doi.org/10.18653/v1/d18-1045.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. Exploiting monolingual data at scale for neural machine translation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference* *on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4205–4215. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1430. URL https://doi.org/10.18653/v1/D19-1430.

Isaac Caswell, Ciprian Chelba, and David Grangier. Tagged back-translation. In Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 1: Research Papers*, pages 53–63. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-5206. URL https://doi.org/10.18653/v1/w19-5206.

Jaehong Park, Jongyoon Song, and Sungroh Yoon. Building a neural machine translation system using only synthetic parallel data. *CoRR*, abs/1704.00253, 2017. URL http://arxiv.org/abs/1704.00253.

Alberto Poncelas, Dimitar Sht. Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. Investigating backtranslation in neural machine translation. *CoRR*, abs/1804.06189, 2018. URL http://arxiv.org/abs/1804.06189.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. Tagged back-translation revisited: Why does it really work? In *ACL*, pages 5990–5997. Association for Computational Linguistics, 2020.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In Toshiaki Nakazawa and Isao Goto, editors, *Proceedings of the 4th Workshop on Asian Translation, WAT@IJCNLP 2017, Taipei, Taiwan, November 27- December 1, 2017*, pages 70–78. Asian Federation of Natural Language Processing, 2017. URL https://www.aclweb.org/anthology/W17-5704/.

Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In Alexandra Birch, Andrew M. Finch, Minh-Thang Luong, Graham Neubig, and Yusuke Oda, editors, *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, NMT@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 18–24. Association for Computational Linguistics, 2018. doi: 10.18653/v1/w18-2703. URL https://doi.org/10.18653/v1/w18-2703.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 820–828, 2016. URL `https://proceedings.neurips.cc/paper/2016/hash/5b69b9cb83065d403869739ae7f0995e-Abstract.html`.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. 2005.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The united nations parallel corpus v1.0. In *LREC*. European Language Resources Association (ELRA), 2016.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz-Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. Paracrawl: Web-scale acquisition of parallel corpora. In *ACL*, pages 4555–4567. Association for Computational Linguistics, 2020.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016a.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016b.

Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *WMT (shared task)*, pages 272–303. Association for Computational Linguistics, 2018.

Raivis Skadins, Karlis Goba, and Valters Sics. Improving SMT for baltic languages with factored models. In *Baltic HLT*, volume 219 of *Frontiers in Artificial Intelligence and Applications*, pages 125–132. IOS Press, 2010.

Matiss Rikters, Marcis Pinnis, and Rihards Krislauks. Training and adapting multilingual NMT for less-resourced and morphologically rich languages. In *LREC*. European Language Resources Association (ELRA), 2018.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* *ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1007. URL `https://www.aclweb.org/anthology/P18-1007/`.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690, 2017.

Rico Sennrich and Barry Haddow. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 83–91. The Association for Computer Linguistics, 2016. doi: 10.18653/v1/w16-2209. URL `https://doi.org/10.18653/v1/w16-2209`.

Andre Tättar, Elizaveta Korotkova, and Mark Fishel. University of tartu's multilingual multi-domain WMT19 news translation shared task submission. In Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 382–385. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-5342. URL `https://doi.org/10.18653/v1/w19-5342`.

Markus Freitag and Yaser Al-Onaizan. Fast domain adaptation for neural machine translation. *CoRR*, abs/1612.06897, 2016. URL `http://arxiv.org/abs/1612.06897`.

Appendix

I. Glossary

MT - Machine translation

NMT - Neural machine translation

RNN - Recurrent neural network

BT - Back translation

WMT - Workshop on machine translation

EN - English language

ET - Estonian language

RU - Russian language

SOTA - State of the art

PDF - Portable document format

HTML - Hypertext markup language

II. Translation output

Table 8. Output from WMT18 EN-ET test set

Forward translation	
Source (EN)	The company, which agreed to sell its stake in Penguin Random House to partner Bertelsmann last month, said its outlook for the year was unchanged after it reported a 1 per cent rise in underlying sales in the first half to 2.05 billion pounds.
Baseline [1]	Ettevõte, kes nõustus müüma oma panuse Penguin Random House partner Bertelsmann eelmisel kuul, ütles, et tema väljavaated aasta oli muutmata pärast seda, kui ta teatas 1 protsendi kasvu aluseks müügi esimesel poolel 2,05 miljardi naela.
BT [2]	Ettevõte, kes nõustus eelmisel kuul müüma oma panuse Penguin Random Maja partnerile Bertelsmannile, ütles, et tema väljavaated aastaks muutusid pärast seda, kui ta teatas esimesel poolajal 1 protsendilist tõusu põhimüügis 2,05 miljardile naelale.
Reference	Ettevõte, mis nõustus müüma eelmisel kuul osaluse ettevõttes Penguin Random House oma partnerettevõttele Bertelsmann, ütles, et nende ootus seoses aastaga ei ole muutunud pärast seda, kui teatati müügitulu 1-protsendilisest kasvust 2,05 miljardi naelani aasta esimeses pooles.
Reversed translation	
Source (EN translationese)	The group with backpacks and dogs moved on to the Viru Keskus crossing to try their luck.
Baseline	Kontserni seljakotid ja koerad liikus Viru Keskuse ristumisse, et proovida oma õnne.
BT	Seljakottide ja koertega koond liikus Viru Keskuse ületamisele, et oma õnne proovida.
Reference	Seltskond kolis seljakottide ja koertega Viru keskuse ristmiku juurde õnne katsuma.

[1]Multilingual baseline described in 5.2
[2]The baseline continued training on 96m of back-translated monolingual data without re-iterating

Table 9. Output from WMT18 ET-EN test set

Forward translation	
Source (ET)	Keskpank ostab turult kokku võlakirju, et innustada võlakirju müünud investoreid raha mujale investeerima.
Baseline	The Central Bank buys bonds from the market to encourage investors who sell bonds to invest money elsewhere.
BT	The central bank is buying bonds from the market together to encourage investors who sold bonds to invest money elsewhere.
Reference	The central bank buys up bonds on the market, to encourage the investors who sold the bonds to invest money elsewhere.
Reversed translation	
Source (ET translationese)	Ajalehe The New York Times 2005. aasta uurimus näitas, et Freeport maksis aastatel 1998 kuni 2004 kohalikele sõjaväelastele ja sõjaväeüksustele ligikaudu 20 miljonit USA dollarit, sealhulgas kuni 150 000 USA dollarit ühele ohvitserile.
Baseline	The 2005 study of The New York Times showed that Freeport paid approximately us $ 20 million to local military and military units between 1998 and 2004, including up to us $ 150 000 to a single officer.
BT	A 2005 study by the newspaper The New York Times revealed that Freeport paid approximately $20 million to local military and military units between 1998 and 2004, including up to $150,000 to one officer.
Reference	A 2005 investigation in The New York Times reported that Freeport paid local military personnel and units nearly $20 million between 1998 and 2004, including up to $150,000 to a single officer.

Table 10. Output from ACCURAT EN-ET test set

1-st example	
Source (EN)	Adequate flow of competent researchers, with high levels of mobility between institutions, disciplines, sectors & countries, is one of the main axes.
Baseline	Pädevate teadlaste piisav voolu, millel on kõrge liikuvus institutsioonide, distsipliinide, sektorite ja riikide vahel, on üks peamisi telge.
BT	Pädevate teadlaste adekvaatne voolu, millel on kõrge liikuvus institutsioonide, distsipliinide, sektorite ja riikide vahel, on üks peamisi telki.
Reference	Piisavalt pädevaid teadlasi, kes on väga liikuvad asutuste, teadusalade, sektorite ja riikide vahel, on üks peamisi telgi.
2-nd example	
Source (EN)	This will be a platform for consumers to experience their favorite commercials or find out more information about a product.
Baseline	See on platvorm tarbijatele kogeda oma lemmik äriühinguid või leida rohkem teavet toote kohta.
BT	See saab olema platvorm, kus tarbijad kogevad oma lemmikkauplejaid või leiavad toote kohta rohkem infot.
Reference	See on klientidele platvorm oma lemmikreklaamide vaatamiseks või toote kohta lisateabe hankimiseks.
3-nd example	
Source (EN)	If a member of the Council declares that, for important and stated reasons of national policy, it intends to oppose the adoption of a decision to be taken by qualified majority, a vote shall not be taken.
Baseline	Kui nõukogu liige kinnitab, et riikliku poliitika olulistel ja märgitud põhjustel kavatseb ta kvalifitseeritud häälteenamusega vastu võtta otsuse vastu, ei tohi hääletada.
BT	Kui nõukogu liige deklareerib , et riikliku poliitika olulistel ja öeldud põhjustel kavatseb ta vastu võtta kvalifitseeritud häälteenamusega langetatud otsuse , ei võta hääletust .
Reference	Kui nõukogu liige teatab, et ta kavatseb liikmesriigi poliitikaga seotud tähtsatel ja esitatud põhjustel olla kvalifitseeritud häälteenamusega otsustamise vastu, siis küsimust hääletusele ei panda.

Table 11. Output from ACCURAT ET-EN test set

1-st example	
Source (ET)	Pärast lühikesi kommentaare viidi need kaks sõjaväehaiglasse tervisekontrolli ja sugulastega kohtuma.
Baseline	After brief comments, these two military hospitals were taken to meet with the health checks and relatives.
BT	After brief comments, these two military hospital checks and relatives were led to meet.
Reference	After making brief comments, the two were taken to a military hospital for medical checks and to reunite with relatives.
2-nd example	
Source (ET)	Meie tulevik sõltub sellest, kas Euroopa saab tõeliselt teadmistepõhiseks ühiskonnaks.
Baseline	Our future depends on whether Europe is truly a knowledge-based society.
BT	Our future depends on whether Europe will truly become a knowledge-based society.
Reference	Our future depends on Europe becoming a true knowledge society.
3-nd example	
Source (ET)	Oma kultuurilise mitmekesisuse kaitsmiseks ja kohalike toodete propageerimiseks taotles ELi Maailma Kaubandusorganisatsioonis niinimetatud kultuurilist erandit, mis õnnestuski saavutada.
Baseline	In order to protect its cultural diversity and promote local products, the EU sought a so-called cultural exception in the World Trade Organization, which was successful.
BT	To protect its cultural diversity and promote local products, the EU sought a so-called cultural exception in the World Trade Organization, which succeeded in achieving.
Reference	To protect its own cultural diversity and promote local productions, the EU sought and secured at the World Trade Organisation what became known as the ‘cultural exception’.

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Dmytro Kolesnykov**,

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,

   Continuous learning for multilingual neural machine translation system,

   supervised by Mark Fišel and 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.

Dmytro Kolesnykov
26/02/2021