Tilde's Machine Translation Systems for WMT 2019

Mārcis Pinnis and **Rihards Krišlauks** and **Matīss Rikters**
Tilde / Vienibas gatve 75A, Riga, Latvia
{firstname.lastname}@tilde.lv

Abstract

The paper describes the development process of Tilde's NMT systems for the WMT 2019 shared task on news translation. We trained systems for the English-Lithuanian and Lithuanian-English translation directions in constrained and unconstrained tracks. We build upon the best methods of the previous year's competition and combine them with recent advancements in the field. We also present a new method to ensure source domain adherence in back-translated data. Our systems achieved a shared first place in human evaluation.

1 Introduction

Since the paradigm-shifting success of neural machine translation (NMT) systems at the 2016 Conference on Machine Translation (WMT) (Bojar et al., 2016), NMT methods and neural network architectures applied in NMT have been annually improved. In 2016, the best-performing systems were based on recurrent neural networks with gated recurrent units (GRU) (Sennrich et al., 2016; Bojar et al., 2016). In 2017, deep GRU models (Sennrich et al.) and models based on shallow multiplicative long short-term memory units (MLSTM; (Pinnis et al., 2017b)) allowed achieving the best results (Bojar et al., a). In 2018, the majority of best-performing systems were based on self-attentional (Vaswani et al., 2017) (Transformer) models (Bojar et al., b).

A year has passed, and the majority of best-performing systems submitted to the shared task on news translation of WMT 2019 are still based on Transformer networks. However, improvements are evident in other areas (e.g., usage of document-level context, very deep models, distillation by ensemble teachers, etc.)[1]. Quite a few of the submissions indicate that substantial amounts of computational resources may have been utilised in order to achieve such results. As we do not have access to large GPU clusters, our strategy for participating at the shared task on news translation of the 2019 Conference on Machine Translation was comprised of combining different methods that showed promising results in scientific publications published in 2018, and analysing whether the methods allowed increasing the overall quality of NMT systems when training NMT models using just modest hardware (with access to one or two graphical processing units) and with the goal of producing models suitable for production.

In our experiments, we investigated methods for corpora filtering (the Tilde MT parallel data filtering (TMTF) and normalisation workflow (Pinnis, 2018) together with dual conditional cross-entropy filtering (DCCEF) (Junczys-Dowmunt, 2018)), training data pre-processing using the methods described by Pinnis et al. (2018a), a new optimisation method, the quasi-hyperbolic Adam, proposed by Ma and Yarats (2018), back-translation with sampling-based decoding (e.g., as done by Edunov et al. (2018)) and by targeting rare words (Fadaee and Monz, 2018) and in-domain subsets of the monolingual data, and automatic linguistically informed post-editing of named entities and non-translatable phrases.

This year, Tilde participated in the shared task on news translation for the English↔Lithuanian language pair. We trained constrained and unconstrained systems for both translation directions.

The paper is further structured as follows: Section 2 describes the data used for training, Section 3 describes the main NMT model training experiments, Section 4 describes our experiments on automatic post-editing of named entities, Section 5 summarises our automatic evaluation results, and the paper is concluded in Section 6.

[1] http://matrix.statmt.org

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 327–334
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

2 Data

Similarly to the year before, we used both constrained data, which were provided by the organisers of the shared task, as well as unconstrained data, which comprised publicly available parallel and monolingual corpora as well as proprietary data from the Tilde Data Library[2]. For language model (LM) training and back-translation, we used news data provided by the organisers. For the unconstrained systems, we used a proprietary news corpus. The raw statistics of data available are provided in Table 1. For validation, we used the first 1000 sentences of the NewsDev2019 data set. Evaluation was performed on NewsTest 2019.

2.1 Data Filtering

For data filtering, we applied the parallel data filtering methods of Tilde MT (Pinnis et al., 2018b; Pinnis, 2018) for both constrained and unconstrained systems. The filters address potential issues that arise from misalignment of parallel data , incomplete translation, various types of data corruption, and other types of data quality issues. However, these filters do not perform data selection. Therefore, we applied also data filtering using DCCEF proposed by Junczys-Dowmunt (2018). As it uses an in-domain LM to discard out-of-domain sentence pairs, it performs the task of data selection. Because for the constrained systems the data-set was not sufficiently large, we applied the filter with a threshold of > 0. For the unconstrained systems, we set the threshold to 11 million[3] highest scored sentence pairs.

For monolingual data, we filtered out all sentences that: 1) were redundant, 2) exceeded 128 tokens or 1000 characters, 3) contained tokens over 50 characters, and 4) contained corrupt characters. See Table 1 for statistics of data filtering.

2.2 Data Pre-Processing

This year, we did not change the parallel and monolingual data pre-processing workflows that we used for our WMT 2018 submissions (Pinnis et al., 2018a).

Similarly to last year, the training corpora were supplemented with synthetic data where up to three words in each sentence were replaced with unknown word identifiers on both source and target sides to ensure that the NMT models are able to handle rare and unknown phenomena during translation (Pinnis et al., 2017a). The statistics of the parallel corpora after supplementing them with synthetic data sets are provided in Table 1.

3 NMT Systems

We took an iterative approach to validating the methods we selected for use in NMT system training. At each step, we either accepted or rejected a method for further use based on its performance compared to a baseline. When moving on, we would often use the previously selected method as a baseline for the next method (which we would combine with the previous method) and so on. More specifically, we conducted the experiments as follows: 1) Filtering (Section 3.1), 2) ~QHAdam (Section 3.2.1), 3) regular back-translation, 4) large batches (Section 3.3), 5.a) back-translation using beam search or sampling (Section 3.4.2), 5.b) back-translation using rare or random data (Section 3.4.1, the results weren't used further), 6) QHAdam (Section 3.2), 7) Source domain adherence (Section 3.4.3), 8) Transformer-big (Section 3.5). The outline of this section loosely follows the above timeline.

As a result of the iterative approach, the evaluation of the training methods was mostly non-exhaustive – meaning that it was usually done only for a single translation direction (most often En → Lt) testing only a few possible configurations (e.g., different model hyper-parameters, back-translated data-set size, etc.). Also, for some experiments we did not methodically test the effect of each of the compounding changes to the experiment's configuration, e.g., in ~QHAdam experiments (in Section 3.2) along with adopting the new optimiser we also selected a new learning-rate and learning-rate schedule without confirming that the baseline optimiser would not also benefit from these changes. As a result, for some experiments we cannot confirm with certainty that the selected method is better than the baseline, only that the selected method with a given set of hyper-parameters is better. The above choices were primarily motivated by resource and time constraints.

All NMT systems described further used the Transformer architecture (Vaswani et al., 2017) and were trained using the Marian NMT toolkit (Junczys-Dowmunt et al., 2018). Unless noted

[2] www.tilde.com/products-and-services/data-library

[3] The threshold was empirically identified by training multiple models with thresholds set at 8 to 12 million.

	Lang. pair	Parallel data (sentence pairs)					Monolingual data (sentences)		
		Raw		+TMTF	+Synth. data	+DCCEF	Raw	Filtered	For LM
		Total	Unique				Total	Unique	Unique
(U)	en-lt	42.9M	30.5M	15.0M	28.6M	11.0M	82.5M	61.3M	4.7M
	lt-en						63.9M	61.0M	4.9M
(C)	en-lt	2.4M	2.3M	1.5M	3.0M	1.7M	103.5M	75.5M	0.7M
	lt-en						63.5M	60.9M	2.0M

Table 1: Training data statistics (TMTF - Tilde MT filtering, DCCEF - dual conditional cross-entropy filtering)

otherwise, we used the *base model* configuration for the model hyper-parameters.

3.1 Filtering

Since DCCEF achieved the best results in the shared task on parallel corpus filtering at WMT 2018 (Koehn et al., 2018), we decided to test whether the combination of our filtering methods (i.e., TMTF) and DCCEF allows acquiring better models. Therefore, we filtered the parallel corpora using DCCEF. For this, we trained two NMT models using the data that were already filtered using TMTF and four language models (two in-domain models that were trained on news corpora and two models trained using the parallel data), and trained several NMT systems. Figure 1 shows the training progress for En → Lt. It is evident that the combination of the methods works better only for the unconstrained systems. We suspect that it is because the unconstrained data sets are large enough to leave enough training data remaining in the filtered data sets. Further, all experiments for unconstrained systems will be performed using data filtered with TMTF and DCCEF and for constrained systems – only TMTF.

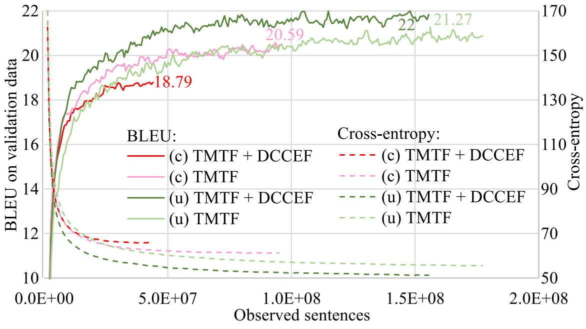

Figure 1: En → Lt systems trained on datasets filtered using the TMTF and DCCEF methods

3.2 QHAdam

We used two versions of the Quasi-Hyperbolic Adam (QHAdam) optimiser (Ma and Yarats, 2018) to train our systems – a version as described in the original paper, and a modified version (∼QHAdam) as described below. The modified version was due to an error in our initial implementation of the optimiser but it performed well enough for us to use it to train the majority of the systems during the period of the competition.

3.2.1 ∼QHAdam

We define the ∼QHAdam's update step in (1). The definitions for g'_t, s'_t, v_1 and v_2 are the same as in the original paper.

The comparison of ∼QHAdam and the baseline system for the constrained En → Lt track is given in Figure 2. ∼QHAdam was tested with different combinations of settings for the learning rate and the number of warm-up steps used. In our initial experiments, we found that setting the learning rate to 5×10^{-4} and using 48k warm-up steps worked best. A workspace size of 9 GB on 2 GPUs was used in Marian which resulted in an effective batch-size of around 255 sentences.

3.3 Using Large Batches

As shown by Popel and Bojar (2018) and Ott et al. (2018), using a large batch size in conjunction with increasing the learning rate allows to train better-performing NMT systems. We confirm these findings. We trained the same system described in Section 3.2.1 except training it with a workspace size of 14 GB on 8 GPUs (simulated using the *--optimizer-delay* option in Marian) which resulted in an effective batch size of ∼1263 sentences. Additionally we increased the learning rate to 7.3×10^{-4} roughly keeping to the rule of scaling the learning rate by a factor of $\sqrt{n}$ when the batch size has increased by a factor of n (Hoffer et al., 2017). The results are given in Figure 3. These experiments were done using back-translated data (see Section 3.4). When using non-back-translated data, we saw overfitting occur.

$$\theta_{t+1} \leftarrow \theta_t - \alpha \left[(1 - v_1) \cdot \nabla \hat{L}_t(\theta_t) + \frac{v_1 \cdot g'_{t+1}}{\sqrt{(1 - v_2)(\nabla \hat{L}_t(\theta_t))^2 + v_2 \cdot s'_{t+1} + \epsilon}} \right] \quad (1)$$

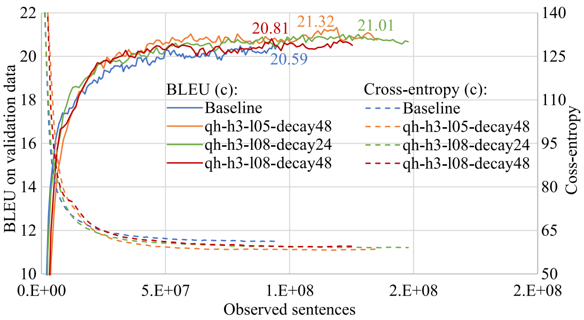

Figure 2: Training progress for the baseline and ~QHAdam systems in the En → Lt translation direction.

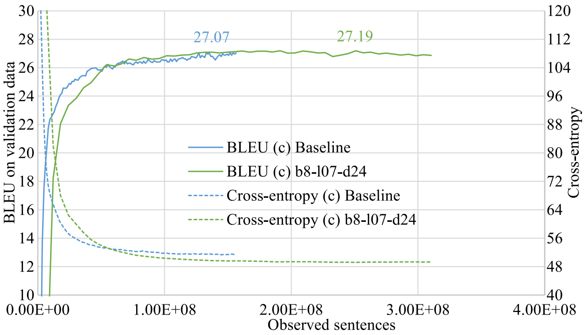

Figure 3: Training progress for ~QHAdam systems comparing effects of different batch sizes and learning rates in the En → Lt translation direction.

3.4 Experiments with Back-translation

We used NMT model adaptation through back-translation (Sennrich et al., 2015) to adapt NMT systems to the news domain. We applied two iterations of back-translation and the subsequent system training to incrementally improve the back-translated data set (Rikters, 2018). We also analysed methods for selection of the data for back-translation. The methods are discussed further. In the figures further, if not specified in the name of each system, the proportion between parallel and back-translated data is 1-to-1.

3.4.1 Rare vs. Random Data for Back-Translation

Fadaee and Monz (2018) showed that adaptation through back-translation works better if the data for back-translation can be considered rare or difficult. Therefore, we compared two types of data selection - random selection and selection by targeting rare words (as proposed by Fadaee and Monz (2018)), back-translated the data sets using beam search, and trained NMT models. Figure 4 depicts the training progress of the En → Lt and Lt → En systems. The results suggest that targeting of sentences containing rare words did not help. We believe that this is due to the fact that what is rare in the target language may not be relevant for speakers of the source language. Therefore, there is no guarantee that the method will work. We stopped here and did not pursue this method further.

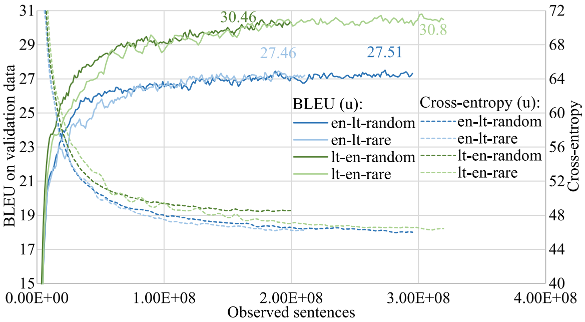

Figure 4: Training progress of systems trained on randomly selected data for back-translation and data selected by targeting rare words

3.4.2 Beam vs. Sampling

As suggested by Edunov et al. (2018), when back-translating data for domain adaptation, better-performing models can be acquired when using sampling instead of beam search. Therefore, we trained several systems on different amounts of back-translated data. The training progress of the systems is depicted in Figure 5.

For the final training iteration, we used sampling instead of beam-search during decoding for all but one system.

3.4.3 Source Domain Adherence

When adapting a system to a specific domain, it is important to use data from that specific domain. However, since we use a monolingual corpus from the target language to adapt an NMT system for source content, there may still be a domain mismatch, because how people write and what they write about in the target language may be (to higher or lower extent) irrelevant for the

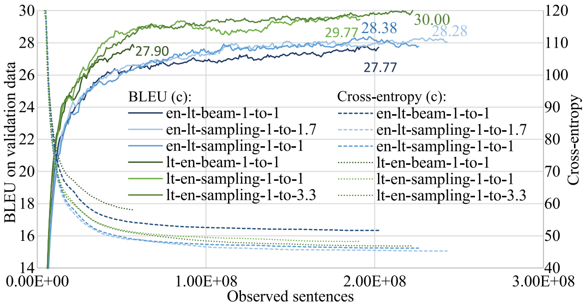

Figure 5: Training progress of systems trained on back-translated data that was acquired using beam search and sampling.

people writing in the source language. Therefore, we performed an experiment where we translated the in-domain source data into the target side using an NMT system, trained a language model on the translations, scored each sentence of the target side's monolingual corpus, and used only the top-scored sentences for back-translation. Comparison between systems trained on randomly selected data for back-translation and data selected using LMs (with *sorted* in the name) is given in Figure 6.

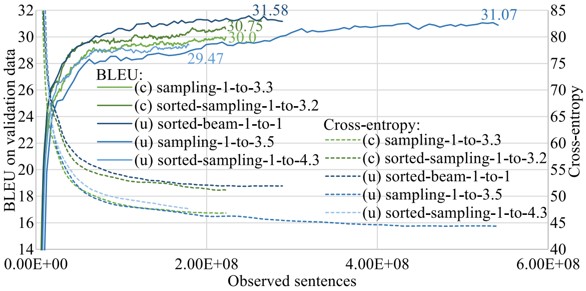

Figure 6: Training progress of systems trained on randomly selected data and data selected using LMs for back-translation.

3.5 Transformer Big

When training the unconstrained systems on the second iteration of back-translated data, we trained a variant for both translation directions using the *transformer-big* configuration (Vaswani et al., 2017). While doing so, we also adjusted the learning rate. Due to time constraints and technical difficulties we were not able to run these experiments to completion. Nonetheless, the *transformer-big* configuration still managed to surpass the baseline. For results see Figure 7.

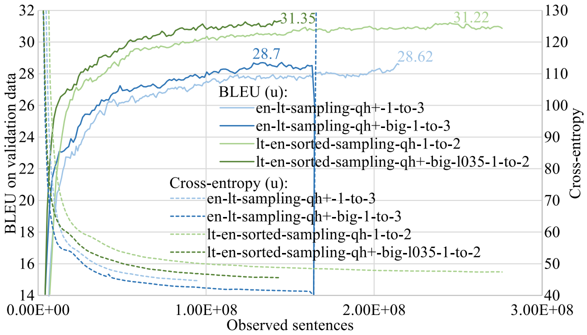

Figure 7: Training progress for the *transformer-big* systems comparing them to QHAdam baselines.

4 Automatic Named Entity Post-Editing

In our submissions for WMT 2018, we introduced an automatic named entity (NE) post-editing (ANEPE) workflow (Pinnis et al., 2018a), which allowed to fix translations of NEs (consisting of one word) and non-translatable words after NMT decoding. The method depends on the quality of word alignments. Because then we did not have methods to extract reliable word alignments from Transformer models, we had to rely on external word alignment using *fast_align* (Dyer et al., 2013). This resulted in many misalignments and unalignments, and incorrect post-edits. This year, we trained all models using the guided alignment method implemented in Marian (Junczys-Dowmunt et al., 2018). Although we still had to pre-process training data using *fast_align*, the NMT models learned to produce more reliable word alignments. We also extended the ANEPE method to support multi-word NEs and non-translatable phrases.

The method works as described further. Using collections of NEs and non-translatable phrases, we perform dictionary-based NE recognition in the source text. Then, for each recognised unit, we analyse whether the NMT translation contains a valid translation of the source unit. In order to support morphologically rich languages (as is Lithuanian), stemming of tokens is performed. However, NEs can already be included in surface forms in the NE collections to account for possible stemming-related issues. If a valid translation is not found, we analyse whether we can identify, which target words the source unit was translated into. If the words are next to each other (i.e., there is no gap between the target words), we replace the target words (except trailing stop-words) with the most similar (according to Levenshtein distance (Levenshtein, 1966)) translation equivalent (except trailing stop-words) found in the NE collection. Stop-words are excluded as the word alignment extracted from the NMT model commonly aligns stop-words to content words when stop-words (dis)appear in the target language. Using ANEPE, we improved the translation quality by 0.04 to 0.1 BLEU points for all submissions. Statistics also show that out of 408 named entities and non-translatable phrases identified in the Lithuanian validation set, 322 already had valid translations, 26 were post-edited, and the remaining 60 either had alignment issues or the target words were too dissimilar from the entries in the NE collection. We applied ANEPE for all our submissions.

5 Results

Automatic evaluation results of our final systems using BLEU[4] (Papineni et al., 2002) are given in Table 2. To acquire final translations, we performed also ensembling of the best-performing individual models. For submission, we selected the best-performing models for both translation directions and both scenarios. However, it is evident that other models were able to translate the NewsTest 2019 evaluation set better (for 3 out of 4 submissions). Although this can be expected, when deciding, which systems to submit, we did not account for the change of the evaluation strategy, i.e., the fact that the evaluation set contained only texts originally written in the source language (which is different from previous years). The results clearly show that the models that are more source domain adherent (e.g., the '(u) so-beam-∼qh-1-to-1' unconstrained system for Lt → En) even surpass the quality of our ensemble models.

System	NewsDev (2019a)	NewsTest (2019)
English-Lithuanian		
(u) best 4 ens.	**27.18**	18.84
(u) best 2 ens.	27.03	**19.53**
(c) best 5 ens.	**26.70**	17.86
(u) sa-∼qh-1-to-3	26.66	18.76
(u) sa-qh+-big-1-to-3	26.61	19.13
(c) best 3 ens.	26.54	**18.59**
(c) sa-qh+-1-to-3.3	26.42	18.14
(c) sa-∼qh-1-to-1.7	26.19	18.17
(c) sa-∼qh-1-to-1	26.16	17.83
Lithuanian-English		
(u) best 5 ens.	**30.41**	31.55
(c) best 5 ens.	**29.76**	**30.21**
(u) so-beam-∼qh-1-to-1	29.43	**31.67**
(u) so-sa-qh+-big-l035-1-to-2	29.12	30.09
(u) so-sa-qh-1-to-2	28.99	29.60
(c) so-sa-∼qh-1-to-3.2	28.84	29.30
(c) so-sa-qh-1-to-3.2	28.66	28.93
(c) sa-1-to-3.3	28.17	28.94

Table 2: Evaluation results - BLEU scores (submitted models are underlined, bold marks best results for both scenarios, (c) - constrained scenario, (u) - unconstrained scenario, 'ens.' - ensembles of models, 'sa' - sampling-based back-translation, 'so' - source domain adherence, 'qh' - quasi-hyperbolic Adam, '∼qh' - modified version of 'qh', 'qh+' - 'qh' with tuned parameters, 'M-to-N' - the proportion of parallel and back-translated data)

6 Conclusion

The paper presented Tilde's efforts on developing NMT systems for the WMT 2019 shared task on news translation. We built upon our methods from the previous year and investigated other novel methods proposed in 2018. Our experiments showed that improvements in translation quality could be achieved by using improved filtering by combining TMTF and DCCEF, sampling-based back-translation (although not for all systems), and the quasi-hyperbolic Adam optimiser. We also introduced a new method that allows to boost the quality of back-translation by ensuring source domain adherence of the data selected for back-translation, as well as described improvements upon our automatic named entity post-editing method. Our systems achieved a shared first place in human evaluation.

Acknowledgements

The research has been supported by the European Regional Development Fund within the research project "Multilingual Artificial Intelligence Based Human Computer Interaction" No. 1.1.1.1/18/A/148. We thank the High Performance Computing Center of Riga Technical University for providing access to their GPU computing infrastructure.

[4] BLEU scores were obtained using SacreBLEU (Post, 2018), checksum: BLEU+case.mixed+numrefs.1 +smooth.exp+tok.13a+version.1.2.7.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. a. findings of the 2017 conference on machine translation (wmt17).

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. b. findings of the 2018 conference on machine translation.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, June, pages 644–648, Atlanta, USA.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Marzieh Fadaee and Christof Monz. 2018. Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446.

Elad Hoffer, Itay Hubara, and Daniel Soudry. 2017. Train Longer, Generalize Better: Closing the Generalization Gap in Large Batch Training of Neural Networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1731–1741. Curran Associates, Inc.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 901–908, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 739–752, Belgium, Brussels. Association for Computational Linguistics.

Vladimir I Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8):707–710.

Jerry Ma and Denis Yarats. 2018. Quasi-Hyperbolic Momentum and Adam for Deep Learning. *arXiv preprint arXiv:1810.06801.*

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling Neural Machine Translation. *arXiv:1806.00187 [cs].* ArXiv: 1806.00187.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Mārcis Pinnis. 2018. Tilde's parallel corpus filtering methods for wmt 2018. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 952–958, Belgium, Brussels. Association for Computational Linguistics.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017a. Neural Machine Translation for Morphologically Rich Languages with Improved Sub-word Units and Synthetic Data. In *Proceedings of the 20th International Conference of Text, Speech and Dialogue (TSD2017)*, volume 10415 LNAI, Prague, Czechia.

Mārcis Pinnis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics. 2017b. Tilde's Machine Translation Systems for WMT 2017. In *Proceedings of the Second Conference on Machine Translation (WMT 2017), Volume 2: Shared Task Papers*, pages 374–381, Copenhagen, Denmark. Association for Computational Linguistics.

Mārcis Pinnis, Matīss Rikters, and Rihards Krišlauks. 2018a. Tilde's machine translation systems for wmt 2018. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 477–485, Belgium, Brussels. Association for Computational Linguistics.

Mārcis Pinnis, Andrejs Vasiļjevs, Rihards Kalniņš, Roberts Rozis, Raivis Skadiņš, and Valters Šics. 2018b. Tilde MT Platform for Developing Client Specific MT Solutions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Martin Popel and Ondej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70. ArXiv: 1804.00247.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Matīss Rikters. 2018. Impact of Corpora Quality on Neural Machine Translation. In *In Proceedings of the 8th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2018)*, Tartu, Estonia.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. the university of edinburgh's neural mt systems for wmt17.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving Neural Machine Translation Models with Monolingual Data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.