Tilde's Machine Translation Systems for WMT 2017

Mārcis Pinnis and **Rihards Krišlauks** and **Toms Miks**
and **Daiga Deksne** and **Valters Šics**
Tilde, Vienibas gatve 75A, Riga, Latvia
{marcis.pinnis,rihards.krislauks,toms.miks}@tilde.lv
{daiga.deksne,valters.sics}@tilde.lv

Abstract

The paper describes Tilde's English-Latvian and Latvian-English machine translation systems for the WMT 2017 shared task in news translation. Both constrained and unconstrained systems are described. Our constrained systems were ranked as the best performing systems according to the automatic evaluation results. The paper gives details to how we pre-processed training data, the NMT system architecture that we used for training the NMT models, the SMT systems and their usage in NMT-SMT hybrid system configurations.

1 Introduction

The year 2016 marked the first time when neural machine translation (NMT) systems achieved significantly better results than statistical machine translation (SMT) systems for most of the translation directions in the news translation shared task of the WMT conference. This was achieved due to a number of architectural and data pre-processing novelties that the winning systems incorporated, for instance, the use of an attention mechanism in the decoder of the NMT system (Bahdanau et al., 2014), back-translation of additional in-domain monolingual data for domain adaptation of the NMT system after training of a broad domain model or during re-training of the whole NMT model, use of sub-word units to address the problem of out-of-vocabulary word translation, and others (Sennrich et al., 2016).

Since then, a number of further advances have been made in machine translation and related fields. A lot of effort has been invested in the search for the best hyper-parameter configurations and neural network architectures for NMT system training (Britz et al., 2017). In particular, the use of long short-term memory (LSTM) cells and deep architectures has shown to allow increasing translation quality. Parallel to that, a number of novelties in neural network architectures have been introduced for other sequence processing tasks, some of which, like the multiplicative LSTM (MLSTM) units (Krause et al., 2016), promise advantages even over deep recurrent network architectures. For data pre-processing, we have shown that the language agnostic word splitting method using byte pair encoding (BPE) inconsistently splits words for morphologically rich languages and that the method can be improved by linguistically motivating word splitting (Pinnis et al., 2017b).

For the WMT 2017 shared task in news translation, we build upon the NMT toolkit Nematus (Sennrich et al., 2016) that achieved the best results in the WMT 2016 shared task. We also incorporate in our systems the latest advancements in the field, for instance, MLSTM recurrent layers, morphology-driven word splitting, better handling of unknown and rare words with robust NMT models, and hybrid methods. The improvements over the baseline NMT model have allowed us to develop the best scoring systems for the English-Latvian and Latvian-English translation directions.

The paper is further structured as follows: Section 2 provides an overview of our WMT 2017 systems, Section 3 describes the data and the different data processing workflows used for preparing the data for training, Section 4 describes SMT systems that were used in NMT-SMT hybrid system configurations, Section 5 describes the NMT architecture used for training of our NMT systems, Section 6 describes the hybrid NMT-SMT system architecture, Section 7 describes our evaluation results, and Section 8 concludes the paper.

2 System Overview

For the WMT 2017 shared task, we developed both constrained and unconstrained MT systems. In total, we submitted five systems:

- Constrained English-Latvian and Latvian-English NMT-SMT hybrid systems.

- Unconstrained English-Latvian and Latvian-English NMT-SMT hybrid systems trained on significantly larger corpora.

- An unconstrained English-Latvian SMT system that achieves higher automatic evaluation results than the NMT-SMT hybrid systems.

3 Data

For training of MT systems, we used the WMT 2017 training data, however, for the unconstrained systems we also used resources from the Tilde Data Library.[1] All data were filtered using our data filtering methods (see Section 3.1), pre-processed with standard and custom pre-processing tools (see Section 3.2), and supplemented with synthetic data (see Section 3.3). For tuning and for decision-making during the development, we used the *newsdev2017* data set that was provided by the WMT 2017 organisers.

3.1 Data Filtering

Our previous research in NMT system development has shown that NMT systems are more sensitive to the noise present in the training data (Pinnis et al., 2017a) than SMT systems, therefore, we performed parallel data filtering to reduce potential non-parallelities and the negative effect of noise on the NMT systems. The filtering consisted of the following steps:

1. Long sentence filtering (longer than 1500 symbols or 80 tokens).

2. Sentence length difference filter (sentence pairs with a length ratio smaller than 0.3 were filtered out).

3. Incorrect language filtering using a language detection tool (Shuyo, 2010).

4. Low content overlap filter using the cross-lingual alignment tool *MPAligner* (Pinnis, 2013).

5. Bad encoding filter that filtered out sentences containing foreign and corrupt symbols.

6. Digit mismatch filter that showed to be an effective method for dealing with sentence segmentation issues in a number of corpora (including the Digital Corpus of European Parliament).

Training data statistics for both constrained and unconstrained systems are shown in Table 1.

3.2 Data Pre-processing

After filtering, all training data were pre-processed using the following steps:

1. *Normalisation of punctuation.* Only one standard of quotation marks and apostrophes were used, hyphenated tokens were split and the hyphens were replaced with a special symbol.

2. *Identification of non-translatable entities.* E-mail addresses, URLs, file addresses and XML tags were identified and replaced with place-holders.

3. *Tokenisation.* For tokenisation, we used the Tilde's regular expression-based tokeniser.

4. *Truecasing.* The Moses *truecase.perl* was used to truecase the first word of each sentence.

5. *Morphology-driven word splitting* (Pinnis et al., 2017b). Tokens were split using a morphological analyser and further processed with byte pair encoding (BPE) (Sennrich et al., 2015) to ensure an open vocabulary. For both languages, we used morphological analysers that were developed by Deksne (2013) using finite state transducer technology.

6. *Factorisation.* Following the work of Sennrich and Haddow (2016), who showed that linguistic input features allow increasing NMT system translation quality, we developed our NMT systems using factored models. Therefore, the source data were further factored using a language-specific tag-

---

Scenario	Lang. pair	Before filtering (Total / Unique)		After filtering (Unique)	
		Parallel	**Monolingual**	**Parallel**	**Monolingual**
Constrained	en-lv	4.51M / 1.92M	38.13M / 28.81M	1.61M	27.75M
	lv-en	4.51M / 1.92M	369.85M / 335.55M	1.61M	330.23M
Unconstrained	en-lv	39.28M / 15.78M	128.28M / 87.60M	12.69M	81.68M
	lv-en	39.28M / 15.78M	416.36M / 360.01M	12.69M	351.99M

Table 1: Training data statistics (sentence counts) for SMT and NMT systems before and after filtering

ger or parser. For Latvian, we used an averaged perceptron-based morpho-syntactic tagger (Nikiforovs, 2014) that was trained on the data from Pinnis and Goba (2011). For English, we used the lexicalized probabilistic parser (Klein et al., 2002) from the *Stanford CoreNLP* toolkit (Manning et al., 2014).

3.3 Synthetic Data

Similarly to the method by Pinnis et al. (2017b) that allows training NMT models that are more robust to unknown and rarely occurring words, we supplemented the parallel training data with synthetic parallel training sentences. To create the synthetic corpus, we performed word alignment on the parallel corpus using *fast-align* (Dyer et al., 2013). Then, we randomly replaced one to three unambiguously (one-to-one) aligned content words with unknown word *<UNK>* placeholders. Finally, we copied factor information from the original factored source sentence to the synthetic sentence.

Using the filtered and the synthetic training data, we trained initial target-to-source NMT models (see Section 5 for details on the NMT architecture). Then, we shuffled the available in-domain monolingual data (news articles or news commentary in the target language) and for each system back-translated a part of the monolingual data from the target language into the source language in order to create additional synthetic source-to-target parallel corpora. The data were selected such that the amount would approximately correspond to the original training data. Experiments with different back-translated data proportions showed that the best results could be achieved with a proportion of 1-to-1.

The back-translated parallel corpora were also supplemented with sentence pairs where content words with unambiguous alignments were randomly replaced with unknown word placeholders. Finally, the additional synthetic data were

	Lang. pair	Synth. *<UNK>* sent.	Re-transl. sent.	Total
(C)	en-lv	1.48M	3.09M	6.19M
	lv-en	1.48M	3.09M	6.19M
(U)	en-lv	11.66M	21.69M	46.04M
	lv-en	11.66M	21.36M	45.71M

Table 2: Synthetic data and final NMT model training data statistics

added to the existing training data. The statistics of the synthetic corpora and the final training data for NMT system training are given in Table 2. It can be seen that the synthetic data creation process increased the size of the training data four times.

4 SMT Systems

SMT systems were trained using Moses (Koehn et al., 2007) in the Tilde MT platform (Vasiļjevs et al., 2012). All systems were trained using the filtered training data (see Table 1). Word alignment was performed using fast-align (Dyer et al., 2013). All SMT systems feature 7-gram translation models and the *wbe-msd-bidirectional-fe-allff*[2] reordering models. The systems have two language models that were trained using KenLM (Heafield, 2011) - an in-domain language model trained on the news article and news commentary corpora and an out-of-domain language model trained on the remaining monolingual data. The systems were tuned using MERT on the *news-dev2017* data set.

5 NMT System Architecture

The NMT system architecture is based on the implementation available with the Nematus toolkit that was used by Sennrich et al. (2016) to produce

---

[2]More about the different types of reordering models in Moses can be found online at http://www.statmt.org/moses/?n=FactoredTraining.BuildReorderingModel

Factor	English	Latvian
Word part	350	360
Position	5	5
Lemma	125	125
Part-of-speech tag	10	-
Syntactic function	10	-
Morpho-syntactic tag	-	10

Table 3: Dimensionality of each factor in the encoder's embedding layer

the top-scoring results for multiple language pairs in the WMT 2016 shared task in news translation. It is an encoder-decoder model with attention. The main distinction of our model is the use of multiplicative long short-term memory cells (Krause et al., 2016) instead of gated recurrent units (GRU) in the encoder and in the first cell of the decoder. We also use linguistic input features as described by Sennrich and Haddow (2016). I.e., each factor of a word part has its own embedding vector and in order to obtain one embedding vector for the whole word part, the individual embedding vectors are concatenated.

In more detail, the encoder's embedding layer has a total of 500 dimensions, which are split among the different input factors as specified in Table 5. It accommodates a vocabulary of 25 thousand sub-word units. The embedding layer is followed by a bidirectional MLSTM layer with 1024 dimensions for gates and cell states.

The decoder has a similar architecture to the implementation in the Nematus toolkit (Sennrich et al., 2017) which improves on the original attention-based NMT model (Bahdanau et al., 2014) by conditioning the attention weights on the previously decoded word in addition to the hidden state at the previous time-step. This is achieved by first computing an intermediate state

$$\hat{s}_t = \text{GRU}(s_{t-1}, e_{y_{t-1}}),$$

then using it to compute the attention context

$$c_t = \text{attention}(\hat{s}_t, h),$$

where s_{t-1} and $e_{y_{t-1}}$ are the decoder's hidden state and the embedding of the decoded word at the previous time-step respectively, and h is the annotation matrix produced by the encoder. The hidden state is then calculated as

$$s_t = \text{GRU}(\hat{s}_t, c_t).$$

We modify this scheme by using an MLSTM cell to calculate the intermediate state

$$(\hat{s}_t, z_t) = \text{MLSTM}(s_{t-1}, z_{t-1}, e_{y_{t-1}}),$$

where $\hat{s}_t$ and z_t are the MLSTM cell's output and hidden states respectively.

Similarly to the encoder, all of the gates and intermediate states of the decoder have a dimensionality of 1024. The decoder's embeddings have a dimensionality of 500.

For training, we also used dropout with the rate of 0.2 for hidden layers, and 0.1 for input and output embedding layers. For optimisation, we used Adadelta (Zeiler, 2012) with a learning rate of 0.0001, and we also used gradient clipping with a threshold of 1.

After training, 5 to 7 models that achieved the highest mixed metric evaluation results on the tuning data (i.e., the *newsdev2017* data set) were selected for ensemble decoding with a beam size of 12.

6 Hybrid System Architecture

After developing the initial NMT models, a preliminary manual analysis of the translations of the English-Latvian constrained system showed that only 34-44% of named entities within the tuning set were translated correctly. At the same time, the SMT system was able to handle approximately 70% of named entities correctly. Taking into account that our previous research in hybrid machine translation system development has shown that SMT systems in hybrid NMT-SMT system scenarios can handle rare and unknown word translation in hybrid scenarios better than the NMT models (Pinnis, 2016) alone, we decided to chain the NMT and SMT systems into a hybrid NMT-SMT system set-up. In the hybrid set-up, a sentence is at first translated with the NMT system, after which rare and unknown words that are left untranslated by the NMT system are translated with the SMT system.

The hybrid translation method performs translation in six steps as follows (see Table 4 for an example of a sentence processed through all of the hybrid translation steps):

1. First, rare and unknown words are identified in the source sentence and replaced by unknown word place-holders. Words are considered rare if they consist of at least one sub-word unit (or a sub-word unit bigram), which

377

Translation step	Example sentence
Source text	`Šodien skatieties` **`Ikaunieces-Admidiņas`** `startu Rio spēlēs.`
Pre-processed text	`šodien skat@@ ieties` **`I@@ kaun@@ iec@@ es - Ad@@ mi@@ di@@ ņas`** `start@@ u Rio spēlē@@ s .`
Text with identified rare words	`šodien skat@@ ieties` β**`ID`**β `-` β**`ID`**β `start@@ u Rio spēlē@@ s .`
NMT translation	`watch the` β**`ID`**β `-` β**`ID`**β `start at the Rio Games today .`
Moses XML with untranslated rare words	`<nmt translation="watch the">šodien skatieties </nmt>`**`Ikaunieces`** `<nmt translation="-">-</nmt>`**`Admidiņas`** `<nmt translation="start at the Rio Games today">šodien startu Rio spēlēs</nmt><nmt translation=".">.</nmt>`
Moses XML with identified untranslated person names	`<nmt translation="watch the">šodien skatieties </nmt><`**`ne translation="Ikauniece" prob="1.0">Ikaunieces</ne>`** `<nmt translation="-">-</nmt><`**`ne translation="Admidina||Admidins" prob="0.95||0.05">Admidiņas</ne>`** `<nmt translation="start at the Rio Games today">šodien startu Rio spēlēs</nmt><nmt translation=".">.</nmt>`
SMT translation	`watch the` **`Ikauniece - Admidina`** `start at the Rio Games today .`
Post-processed translation	`Watch the` **`Ikauniece-Admidina`** `start at the Rio Games today.`
NMT only transl. (for comparison)	`Today, look at the start of the` **`Isolence-Admidias`** `in the Rio Games.`

Table 4: Example of the NMT-SMT hybrid translation process

	Lang. pair	BPE count	BPE 2-gram
(C)	en-lv	25	1
	lv-en	35	0
(U)	en-lv	100	1
	lv-en	125	1

Table 5: Rare word detection thresholds

occurrence count in the training data is below a certain threshold. The thresholds for our submitted systems were empirically identified by analysing the hybrid method's performance on the tuning data. The thresholds are given in Table 5.

2. Then, the pre-processed sentence is translated with the NMT system. Our NMT models have been trained to leave the unknown word place-holders untranslated, i.e., to pass them through to the target side (Pinnis et al., 2017b). The capability of the NMT system to pass the place-holders through unchanged is vital for the further steps to work.

3. After translation, the NMT model's produced attention matrix is used to perform word alignment. Here, we also identify which source words correspond to each place-holder on the target side.

4. Then, a Moses XML document is prepared for the sentence such that the Moses SMT system will have to translate only the words that were replaced by the place-holders but leave the remaining part as it was translated by the NMT system.

5. Then, for the Latvian-English unconstrained system, we use a person name and surname dictionary to look-up translations of untranslated person names. The translations from the dictionary are merged in the Moses XML document so that the SMT system would be constrained to the translations found in the dictionary.

6. Finally, the Moses XML document is translated with the SMT system.

In the hybrid set-up, the same pre-processing and post-processing steps are used as for the individual NMT and SMT systems.

7 Results

We evaluated all MT systems using multiple automatic evaluation metrics including BLEU (Papineni et al., 2002), BEER 2.0 (Stanojevic and Sima'an, 2014), CharacTER (Wang et al., 2016),

Scenario	Lang. pair	System	BLEU (CS)	BEER 2.0	CharacTER	TER (CS)
(C)	en-lv	SMT	12.98 (12.36-13.60)	0.5086	0.6642	0.7582
		NMT	†19.49 (18.71-20.28)	0.5478	0.5877	0.6741
		Hybrid	†19.52 (18.70-20.34)	0.5482	0.5853	0.6729
	lv-en	SMT	15.47 (14.88-16.06)	0.5219	0.6606	0.7272
		NMT	†20.01 (19.31-20.64)	0.5494	0.6088	0.6725
		Hybrid	†20.06 (19.45-20.71)	0.5496	0.6081	0.6721
(U)	en-lv	SMT	20.43 (19.57-21.28)	0.5491	0.6126	0.6954
		NMT	20.04 (19.22-20.78)	0.5563	0.5832	0.6634
		Hybrid	20.08 (19.30-20.85)	0.5567	0.5827	0.6630
	lv-en	SMT	19.05 (18.42-19.67)	0.5515	0.6233	0.6928
		NMT	†22.02 (21.38-22.63)	0.5677	0.5838	0.6450
		Hybrid	†22.06 (21.41-22.74)	0.5683	0.5833	0.6442

Table 6: Automatic evaluation results of Tilde's systems (CS stands for case sensitive evaluation; the results are significant compared to the SMT system with $p = 0.01$†; the BLEU scores are given with a 95% confidence interval that was calculated using bootstrap resampling (Koehn, 2004))

and TER (Snover et al., 2006). The automatic evaluation results (see Table 6) on the *new-stest2017* data set show that for English-Latvian the constrained NMT system and for Latvian-English both the constrained and unconstrained NMT systems achieve significantly better results than the SMT systems. The difference between the quality of the unconstrained English-Latvian SMT and NMT systems is not statistically significant.

Since the automatic metrics have shown not to be sufficient to evaluate MT systems of the two different paradigms (Pinnis et al., 2017a), we also performed (blind) human comparative evaluation of the SMT and NMT system translations. Five professional translators were given the source sentence and translations of two MT systems and asked to select, which system (NMT, SMT, or neither) produces a better translation. The evaluation was performed on the tuning data set. In total, 200-250 sentences were evaluated in each evaluation task. The results in Figure 1 show that the NMT system translations are preferred more than the translations of the SMT systems. According to the methodology by Skadiņš et al. (2010), the results are weakly sufficient for all scenarios (except the Latvian-English unconstrained scenario for which the results are strongly sufficient) to conclude that the NMT systems produce better translations than the SMT systems.

The results also show that there is an insignificant quality increase for the hybrid systems over the NMT systems. The increase is minimal as only sentences that contain words with rare word parts

are translated differently. However, the hybrid scenario (and the components used in the hybrid scenario) allows us to integrate the NMT systems in our existing SMT infrastructure for formatting-rich document translation, which is a vital requirement for us to provide NMT services for customers.

Compared to other submitted systems, it is evident (see Table 7) that our constrained NMT-SMT hybrid systems significantly outperform other submitted systems.

8 Conclusions

In the paper, we have described English-Latvian and Latvian-English MT systems that were developed by Tilde for the WMT 2017 shared task in news translation. In total, we submitted five systems: four NMT-SMT hybrid systems (two constrained and two unconstrained systems) and one unconstrained English-Latvian SMT system that achieves similar translation quality as the NMT system according to automatic evaluation.

We have documented the methodology used to prepare the data for training of the systems, the SMT and NMT system training set-ups, the workflow for chaining the NMT and SMT systems into a hybrid NMT-SMT system, as well as our evaluation efforts.

The automatic and manual evaluation results show that three out of four NMT systems significantly outperform the SMT systems. Although the hybrid systems did not produce a significant improvement, the minimal improvement is con-
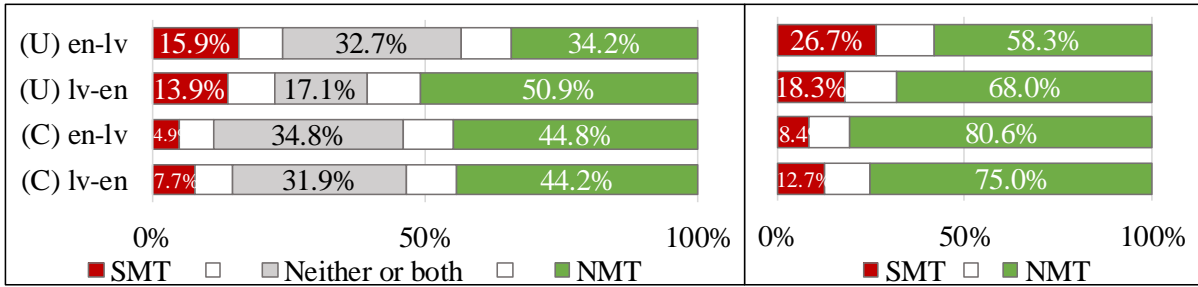
Figure 1: Results of the human comparative evaluation of Tilde's SMT and NMT systems

Lang. pair	System	BLEU (CS)	BEER 2.0	CharacTER	TER (CS)
en-lv	Tilde (hybrid)	†19.52 (18.74-20.31)	0.5482	0.5853	0.6729
	QT21 combination	18.03 (17.32-18.73)	0.5403	0.6455	0.7034
	KIT primary	17.72 (17.01-18.39)	0.5428	0.6051	0.6992
lv-en	Tilde (hybrid)	†20.83 (20.13-21.49)	0.5496	0.6081	0.6641
	UEDIN NMT	20.02 (19.39-20.63)	0.5462	0.6308	0.6719
	JHU SMT	17.67 (17.11-18.30)	0.5281	0.6485	0.7068

Table 7: Automatic evaluation results of the top three English-Latvian and Latvian-English constrained systems submitted for the WMT 2017 shared task on news translation (CS stands for case sensitive evaluation; the results are significant compared to other systems with $p = 0.01$†; the BLEU scores are given with a 95% confidence interval that was calculated using bootstrap resampling (Koehn, 2004))

sistent across all language pairs. The results also showed that in terms of automatic evaluation our submitted NMT-SMT hybrid systems significantly outperform the systems submitted by other participants of the shared task.

Acknowledgments

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Denny Britz, Anna Goldie, Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906* .

Daiga Deksne. 2013. Finite State Morphology Tool for Latvian. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*. pages 49–53.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of NAACL HLT 2013*. Atlanta, USA, June, pages 644–648.

Kenneth Heafield. 2011. KenLM : Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2009, pages 187–197.

Dan Klein, Christopher D Manning, et al. 2002. Fast exact inference with a factored model for natural language parsing. *Advances in Neural Information Processing Systems (NIPS 2002)* pages 3–10.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. pages 388–395.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 177–180.

Ben Krause, Liang Lu, Iain Murray, and Steve Renals. 2016. Multiplicative lstm for sequence modelling. *arXiv preprint arXiv:1609.07959* .

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David Mc-Closky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the ACL 2014 System Demonstrations*. pages 55–60.

Peteris Nikiforovs. 2014. Latvian NLP: Perceptron Tagger. https://github.com/pdonald/latvian.

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*. Association for Computational Linguistics, pages 311–318.

Mārcis Pinnis. 2013. Context Independent Term Mapper for European Languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013)*. Hissar, Bulgaria, pages 562–570.

Mārcis Pinnis and Kārlis Goba. 2011. Maximum Entropy Model for Disambiguation of Rich Morphological Tags. In Cerstin Mahlow and Michael Piotrowski, editors, *Proceedings of the 2nd International Workshop on Systems and Frameworks for Computational Morphology*. Springer Berlin Heidelberg, Zurich, Switzerland, pages 14–22.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017a. Evaluation of Neural Machine Translation for Highly Inflected and Small Languages. In *Proceedings of the 18th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2017)*. Budapest, Hungary.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017b. 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)*. Prague, Czechia.

Mrcis Pinnis. 2016. Towards Hybrid Neural Machine Translation for English-Latvian. In *Human Language Technologies The Baltic Perspective - Proceedings of the Seventh International Conference Baltic HLT 2016*. IOS Press, Riga, Latvia, pages 84–91.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357* .

Rico Sennrich and Barry Haddow. 2016. Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation (WMT 2016) - Volume 1: Research Papers*. pages 83–91.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2015)*. Association for Computational Linguistics, Berlin, Germany.

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 (WMT 2016), Volume 2: Shared Task Papers*.

Nakatani Shuyo. 2010. Language detection library for java. http://code.google.com/p/language-detection/.

Raivis Skadiņš, Kārlis Goba, and Valters Šics. 2010. Improving SMT for Baltic Languages with Factored Models. In *Human Language Technologies: The Baltic Perspective: Proceedings of the Fourth International Conference, Baltic HLT 2010*. IOS Press, volume 219, pages 125–132.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas*. Cambridge, MA, USA, August, pages 223–231.

Miloš Stanojevic and Khalil Sima'an. 2014. B EER : BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. pages 414–419.

Andrejs Vasiļjevs, Raivis Skadiņš, and Jörg Tiedemann. 2012. LetsMT!: A Cloud-Based Platform for Do-It-Yourself Machine Translation. In Min Zhang, editor, *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, Jeju Island, Korea, July, pages 43–48.

Weiyue Wang, Jan-thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTER : Translation Edit Rate on Character Level. In *Proceedings of the First Conference on Machine Translation (WMT 2016), Volume 2: Shared Task Papers*. Berlin, Germany, volume 2, pages 505–510.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .