Tilde's Machine Translation Systems for WMT 2018

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

Abstract

The paper describes the development process of the Tilde's NMT systems that were submitted for the WMT 2018 shared task on news translation. We describe the data filtering and pre-processing workflows, the NMT system training architectures, and automatic evaluation results. For the WMT 2018 shared task, we submitted seven systems (both constrained and unconstrained) for English-Estonian and Estonian-English translation directions. The submitted systems were trained using Transformer models.

1 Introduction

Neural machine translation (NMT) is a rapidly changing research area. Since 2016 when NMT systems first showed to achieve significantly better results than statistical machine translation (SMT) systems (Bojar et al., 2016), the dominant neural network (NN) architectures for NMT have changed on a yearly (and even more frequent) basis. The state-of-the-art in 2016 were shallow attention-based recurrent neural networks (RNN) with gated recurrent units (GRU) (Sennrich et al., 2016) in recurrent layers. In 2017 (Bojar et al., 2017), multiplicative long short-term memory (MLSTM) units (Pinnis et al., 2017c) and deep GRU (Sennrich et al., 2017a) models were introduced in NMT. The same year, self-attentional (Transformer) models were introduced (Vaswani et al., 2017). Consequently, in 2018, most of the top scoring systems in the shared task on news translation of the Third Conference on Machine Translation (WMT) were trained using Transformer models[1]. However, it is already evident that the state-of-the-art architectures will be pushed even further in 2018 (beyond WMT 2018). For instance, Chen et al. (2018) have recently proposed RNMT+ models that combine deep LSTM-based models with multi-head attention and showed that the models outperform Transformer models.

In WMT 2017, Tilde participated with MLSTM-based NMT systems (Pinnis et al., 2017c). In this paper, we compare the MLSTM-based models with Transformer models for English-Estonian and Estonian-English and we show that the state-of-the-art of WMT 2017 is well behind the new models. Therefore, for WMT 2018, Tilde submitted NMT systems that were trained using Transformer models.

The paper is further structured as follows: Section 2 provides an overview of systems submitted for the WMT 2018 shared task on news translation, Section 3 describes the data used to train the NMT systems and the data pre-processing workflows, Section 4 describes all NMT systems trained and experiments on handling of named entities and combination of systems, Section 5 provides automatic evaluation results, and Section 6 concludes the paper.

2 System Overview

For the WMT 2018 shared task on news translation, Tilde submitted both constrained and unconstrained NMT systems (7 in total). The following is a list of the five MT systems submitted:

- Constrained English-Estonian and Estonian-English NMT systems (*tilde-c-nmt*) that were deployed as ensembles of averaged factored data (see Section 3) Transformer models. The models were trained using parallel data and back-translated data in a 1-to-1 proportion.
- Unconstrained English-Estonian and Estonian-English NMT systems (*tilde-*

[1] All 14 of the best automatically scored systems according to the information provided by participants in the official submission portal http://matrix.statmt.org were indicated as being based on Transformer models.

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 477–485

https://doi.org/10.18653/v1/W18-6423

nc-nmt) that were deployed as averaged Transformer models. These models were also trained using back-translated data similarly to the constrained systems, however, the data, taking into account their relatively large size, were not factored.

- A constrained Estonian-English NMT system (*tilde-c-nmt-comb*) that is a system combination of six factored data NMT systems.

- Constrained English-Estonian and Estonian-English NMT systems (*tilde-c-nmt-2bt*) averaged from multiple best NMT models. The models were trained using two sets of back-translated data in a 1-to-1 proportion to the clean parallel data – one set was back-translated using a system trained on parallel-only data and the other set – using an NMT system trained on parallel data and the first set of back-translated data.

3 Data

Data preparation was done using one of two distinct workflows – we used the full workflow for *tilde-c-nmt*, *tilde-nc-nmt* and *tilde-c-nmt-comb* submissions. For the *tilde-c-nmt-2bt* submission we used the light data preparation workflow.

3.1 Full Workflow

For training of the constrained systems, only data provided by the WMT 2018 organisers were used, however, for training of the unconstrained systems, we also used other publicly available and proprietary corpora that were available in the Tilde Data Library[2]. All parallel corpora were filtered (see Section 3.1.1), pre-processed (see Section 3.1.2), and supplemented with additional generated data (see Section 3.1.3).

3.1.1 Data Filtering

As NMT systems are sensitive to noise in parallel data (Pinnis et al., 2017a), all parallel data were filtered using the parallel data filtering methods described by Pinnis (2018). The parallel corpora filtering methods remove sentence pairs that have indications of data corruption or low parallelity (e.g., source-target length ratio, content overlap, digit mismatch, language adherence, etc.) issues. Contrary to Tilde's submissions for WMT 2017, isolated sentence pair filtering for the WMT 2018 submissions was supplemented with a maximum content overlap filter (i.e. only one target sentence for each source sentence was preserved and vice versa based on the content overlap filter's score for each sentence pair).

For filtering, we required probabilistic dictionaries, which were obtained from the parallel corpora (different dictionaries for the constrained and unconstrained scenarios) using *fast_align* (Dyer et al., 2013). The dictionaries were filtered using the transliteration-based probabilistic dictionary filtering method by Aker et al. (2014).

During filtering, we identified that one of the corpora that were provided by the organisers contained a significant amount of data corruption. It was the Estonian↔English ParaCrawl corpus[3]. The corpus consisted of 1.30 million sentence pairs out of which 0.77 million were identified as being corrupt. To reduce the high level of noise, this corpus was filtered using stricter content overlap (a threshold of 0.3 instead of 0.1) and language adherence filters (both the language detection and the valid alphabet filters had to validate a sentence pair instead of just one of the filters) than all other corpora. As a result, only 0.17 million sentence pairs from the ParaCrawl corpus were used for training of the constrained systems. Due to the quality concerns, the corpus was not used for training of the unconstrained systems.

The corpora statistics before and after filtering are provided in Table 1.

3.1.2 Data Pre-processing

All corpora were pre-processed using the parallel data pre-processing workflow from the Tilde MT platform (Pinnis et al., 2018) that performs the following pre-processing steps:

- First, parallel corpora are cleaned by removing HTML and XML tags, decoding escaped symbols, normalising whitespaces and punctuation marks, replacing control characters with spaces, etc. This step is performed only on the training data.

- Then, non-translatable entities, such as e-mail addresses, URLs, file paths, etc. are identified and replaced with place-holders. This allows reducing data sparsity where it is not needed.

[2]Tilde Data Library is an integral component of the Tilde MT platform that provides access to parallel and monolingual data for MT system development (http://www.tilde.com/mt/).

[3]https://paracrawl.eu/download.html

Workflow	Scenario	Before filtering (Total / Unique)	After filtering (Unique)
Full	(C)	2,178,025 / 1,932,954	968,232
	(U)	75,215,347 / 24,660,087	18,755,230
Light	(C)	2,178,025	998,679

Table 1: Training data statistics (sentence counts) before and after filtering

- Then, the data are tokenised using the Tilde MT regular expression-based tokeniser.
- The Moses (Koehn et al., 2007) truecasing script *truecase.perl* is used to truecase the first word of every sentence.
- Then, tokens are split into sub-word units (Sennrich et al., 2015) using byte-pair encoding (BPE) (Gage, 1994). For the constrained and unconstrained systems, we use BPE models consisting of 24,500 and 49,500 merging operations respectively.
- Finally, data for the constrained systems are factored using an averaged perceptron-based morpho-syntactic tagger (Nikiforovs, 2014) for Estonian and the lexicalized probabilistic parser (Klein et al., 2002) from the *Stanford CoreNLP* toolkit (Manning et al., 2014) for English. Similarly to Sennrich and Haddow (2016), we introduce also a factor indicating a word part's position in a word (beginning, middle, end, or the word part represents the whole word - *B*, *I*, *E*, or *O*). As a result, the Estonian data consist of the the following factors: *word part*, *position*, *lemma*, and *morpho-syntactic tag*. The English data consist of the following factors: *word part*, *position*, *lemma*, *part-of-speech tag*, and *syntactic function*.

3.1.3 Synthetic Data

Similarly to Tilde's 2017 systems (Pinnis et al., 2017c), we submitted systems that were trained using synthetic data: 1) back-translated data, and 2) data infused with unknown token identifiers. The back-translated data allow performing domain adaptation and the second type of synthetic data allow training NMT models that are robust to unknown phenomena (e.g., code-mixed content, target language words in the source text, rare or unseen words, etc.) (Pinnis et al., 2017b).

To create the synthetic corpora with unknown phenomena, we extracted *fast_align* (Dyer et al., 2013) word alignments for each sentence pair in the parallel corpora and randomly replaced one to three unambiguously (one-to-one) aligned content words with unknown word identifiers. These synthetic corpora were added to the parallel corpora, thereby almost doubling the sizes of the available training data.

	Lang. pair	Back-transl. sent.	Synth. <*UNK*> sent.	Total
Full workflow				
(C)	en-et	0.97M	1.72M	3.65M
	et-en	0.97M	1.79M	3.73M
(U)	en-et	16.21M	28.10M	63.07M
	et-en	18.39M	30.77M	67.91M
Light workflow				
(C)	en-et	2.11M		3.11M
	et-en	2.05M		3.04M

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

The back-translated data were acquired from two sources: 1) the constrained system data were acquired from initial Transformer-based NMT systems that were trained on the filtered and pre-processed parallel data, which were supplemented with the unknown phenomena infused data, and 2) the unconstrained system data were acquired from pre-existing unconstrained MLSTM-based NMT systems – the NMT systems that were developed by Tilde for the Estonian EU Council Presidency in 2017 (Pinnis and Kalniņš, 2018). In order to limit noise, the back-translated data were filtered using the same parallel data filtering methods that were described in Section 3.1.1 (although with a higher threshold for the content overlap filter). Furthermore, in order to train the final systems, we also generated unknown phenomena infused data for the back-translated filtered data, thereby also almost doubling the sizes of the back-translated data.

The synthetic corpora statistics and the sizes of the total training data are given in Table 2.

	Name	Model	Voc.	Emb. layer (f1:..:fN)	Other layers (enc:dec, size)	Seq. len.
English-Estonian						
(C)	MLSTM	MLSTM	25k	350:5:125:10:10	1:1 1024	80
	transf	Transformer		512:5:125:11:11	6:6, model: 512	128
	transf-2bt		50k	512		
	transf-l		25k	720:5:125:11:11	7:7, model: 720	
(U)	transf-u		50k	512		
Estonian-English						
(C)	MLSTM	MLSTM	25k	360:5:125:10	1:1 1024	80
	transf	Transformer		512:5:125:14	6:6, model: 512	128
	transf-l			720:5:125:14	7:7, model: 720	
	transf-l2			1024:5:125:14	8:8, model: 1024	
	transf-2bt		50k	512	6:6, model: 512	
(U)	transf-u			720		

Table 3: NMT system training configuration (all other parameters were set to the default values of the respective toolkits (Nematus or Sockeye)

3.2 Light Workflow

In the light workflow we used data cleaning and pre-processing methods described by Rikters (2018). The filtering part includes the following filters: 1) unique parallel sentence filter; 2) equal source-target filter; 3) multiple sources - one target and multiple targets - one source filters; 4) non-alphabetical filters; 5) repeating token filter; and 6) correct language filter. The pre-processing consists of the standard Moses (Koehn et al., 2007) scripts for tokenising, cleaning, truecasing, and Subword NMT for splitting into subword units. The filters were applied to the given parallel sentences, monolingual news sentences before performing back-translation, and both sets of synthetic parallel sentences that resulted from back-translating the monolingual news.

4 NMT Systems

In order to train the NMT systems, we used the Nematus (Sennrich et al., 2017b) (for MLSTM models) and Sockeye (Hieber et al., 2017) (for Transformer models) toolkits. All models were trained until convergence (i.e., until an early stopping criterion was met).

4.1 Full Workflow

First, we trained constrained system baseline models using the filtered datasets. For baseline models, we used the *MLSTM* and *transf* configurations (see Table 3). Then, we used the best-performing models (based on translation quality on the validation set), which were the Transformer models (see Figure 1), and back-translated monolingual data. As mentioned before, for the unconstrained systems, we back-translated the monolingual data using pre-existing MLSTM-based NMT systems. Then, using the final training data (parallel and the two synthetic corpora), we trained final Transformer models. For the constrained scenario, we trained multiple models (three for each translation direction) by experimenting with multiple model configurations. For the unconstrained scenario, we trained one model in each of the directions.

In order to acquire the translations for the submissions, we performed model averaging and ensembling as follows:

- For the *tilde-c-nmt* (constrained NMT) systems, we performed model averaging of the best four models (according to perplexity) of the three different run NMT systems and deployed the averaged models in an ensemble.
- For the *tilde-nc-nmt* (unconstrained NMT) systems, we performed model averaging of the best four models.
- For the *tilde-c-nmt-comb* Estonian-English system, we performed majority voting (see Section 4.3) of translations produced by six different runs of different constrained systems (using best BLEU (Papineni et al., 2002) models, averaged models, ensembled averaged models, ensembled models, and larger beam search (10 instead of 5)).

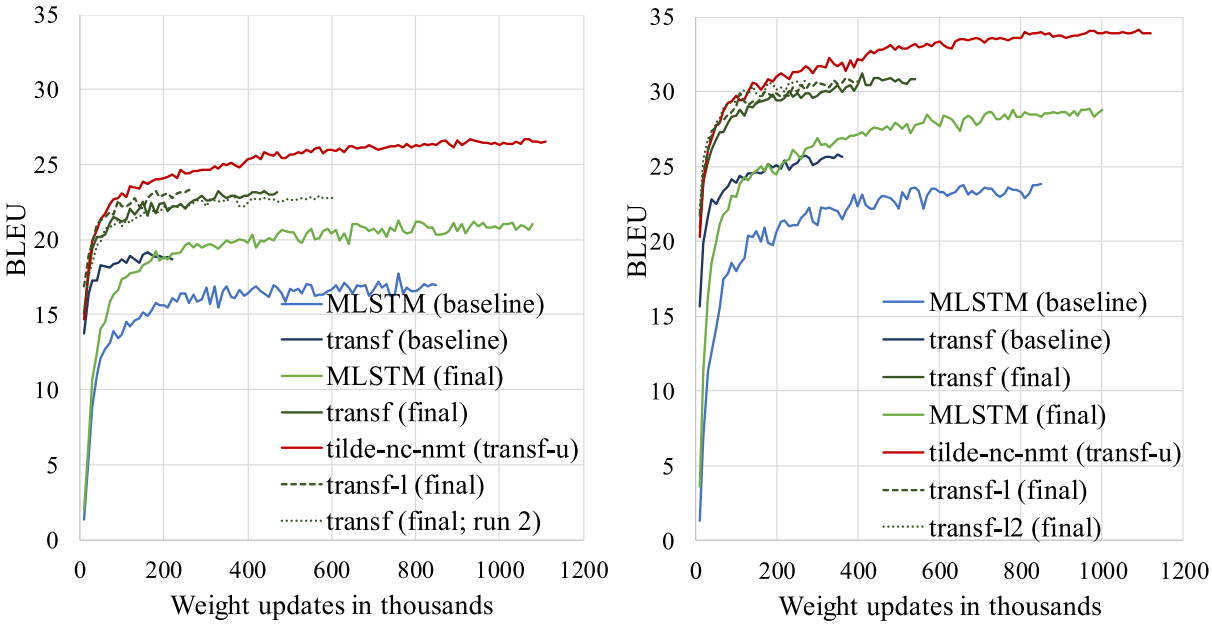

Figure 1: NMT system training progress (BLEU scores on the validation set) for English-Estonian (left) and Estonian-English (right). Note that batch size may differ between different architectures and BLEU scores are calculated on raw (token level) pre-processed validation sets, therefore, the scores are slightly higher than evaluation results for the final translations!

4.1.1 Automatic Post-editing of Named Entities

NMT models so far have struggled with translating rare or unseen words (not different surface forms, but rather different words) correctly (Pinnis et al., 2017c). Named entities and non-translatable entities (various product names, identifiers, etc.) are often rare or unknown. In order to aid the NMT model in translating such tokens better, we extracted named entity and non-translatable token dictionaries from the parallel corpora. This was done by performing word alignment of the parallel corpora using *fast_align* (Dyer et al., 2013) and searching (in a language-agnostic manner) for transliterated source-target word pairs using a similarity metric based on Levenshtein distance (Levenshtein, 1966), which start with upper-case letters. The dictionaries consist of 15.6 (94.7) thousand and 6.2 (149.8) thousand entries for the constrained (unconstrained) English-Estonian and Estonian-English NMT systems respectively.

When the NMT systems had translated a sentence, source-to-target word alignment was extracted from the source sentence and the translation. Then named entity recognition (based on dictionary look-up) was performed on the source text and, if a named entity was found, the target translation was validated against the entries in the dictionary. In order to capture different surface forms, a stemming tool was used. If a translation was contradicting the entries in the dictionary, it was replaced with the closest matching (by looking for the longest matching suffix) translation from the dictionary.

The automatic post-editing method for named entities has a marginal impact on translation quality, however, manual analysis showed that more named entities were corrected than ruined.

4.2 Light Workflow

The light workflow was used to produce the *tilde-c-nmt-2bt* (constrained NMT with two sets of back-translated data) systems. First, we trained baseline models using only filtered parallel datasets (Parallel-only in Figure 2). Then, we back-translated the first batches of monolingual news data and trained intermediate NMT systems (Parallel + First Back-translated). Finally, we used the intermediate NMT systems to back-translate the second batches of monolingual news data and trained final NMT systems (Parallel + Second Back-translated). The training progress in Figure 2 shows that the English-Estonian system benefits from the additional data, but the system in the other direction – not so much.

For the final translations, we used a post-processing script (Rikters et al., 2017) to replace

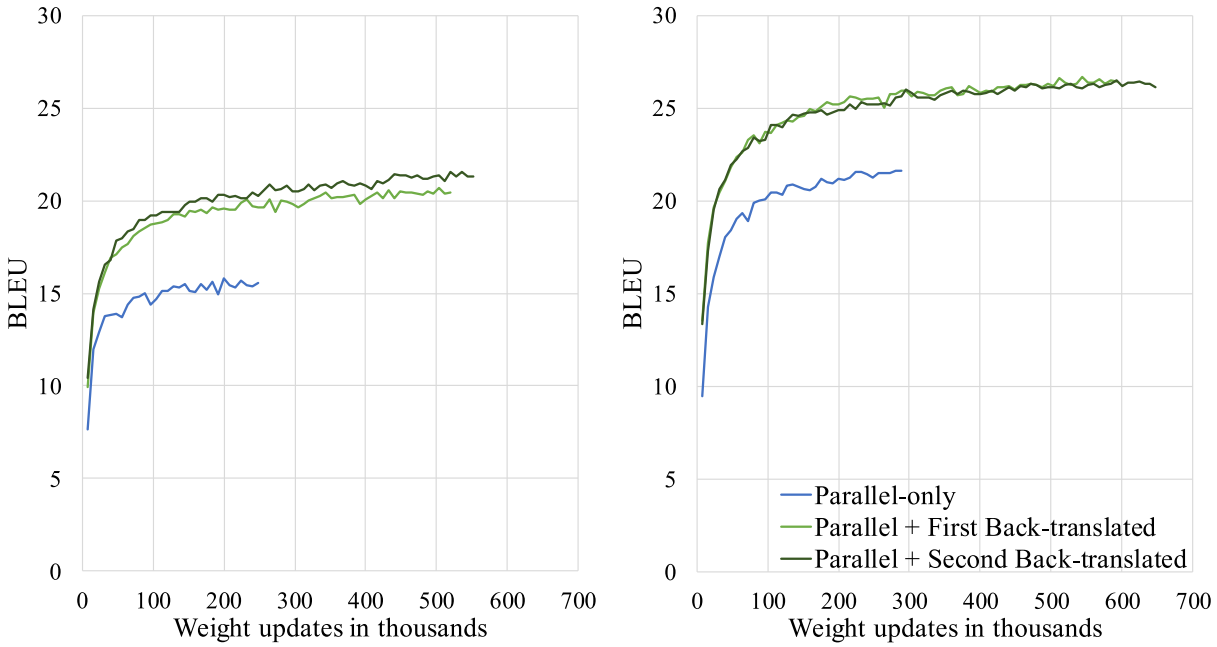

Figure 2: NMT system training progress (SacreBLEU scores on the validation set) for English-Estonian (left) and Estonian-English (right).

consecutive repeating n-grams and repeating n-grams that have a preposition between them (i.e., *victim of the victim*) with a single n-gram. This problem was more apparent in RNN-based NMT systems, but it was also noticable in our Transformer model outputs.

4.3 System Combination

We attempted to increase the quality of existing translations by employing a voting scheme in which multiple machine translation outputs are combined to produce a single translation. We used a custom implementation of the majority voting algorithm (Freitag et al., 2014) to combine six of our best-scoring outputs in the Estonian-English translation direction in the constrained scenario. We did not perform the combination for English-Estonian due to lack of support for alignment extraction for Estonian in Meteor (Denkowski and Lavie, 2014).

MT system translation combination happens on the sentence level. The majority voting scheme assumes a single base translation hypothesis (primary hypothesis) which is aligned at the word level to each of the other hypotheses (secondary hypotheses). The alignments are used to generate a table of all possible word translations relative to each position in the primary hypothesis. The table is then used to count the number of occurrences of different translations. The word translations with the highest count at each position constitute the resulting combined hypothesis.

To acquire the necessary word alignments we used Meteor. Meteor outputs were then converted to a more easily manageable form using the Jane toolkit (Freitag et al., 2014) (we used an *awk* script distributed with Jane). The majority voting algorithm was implemented in *Python*.

5 Results

We performed automatic evaluation of the NMT systems using the SacreBLEU evaluation tool (Post, 2018). The results (see Table 4) show that the Transformer models achieved better results than the MLSTM-based models. For the constrained scenarios, both ensembles of averaged models achieved higher scores than each individual averaged model. It is also evident that the unconstrained models (*tilde-nc-nmt*) achieved the best results.

Although the unconstrained models were not trained on factored data, the datasets were 17 times larger than the constrained datasets. However, the difference is rather minimal and shows that the current NMT architectures may not able to learn effectively from large datasets.

The official human evaluation results (see Table 5) from the WMT 2018 shared task on news translation (Bojar et al., 2018) show that

System	Configuration	BLEU
English-Estonian		
MLSTM (final)	5 model ensemble	20.80
transf (final)	4 model average	22.82
transf-l (final)		23.04
transf (final; run 2)		22.56
tilde-c-nmt	ensemble of 3 averaged models	**23.54**
tilde-c-nmt-2bt	3 model average	**23.57**
tilde-nc-nmt (transf-u)	4 model average	**24.35**
Estonian-English		
MLSTM (final)	5 model ensemble	26.79
transf (final)	4 model average	28.14
transf-l (final)		28.83
transf-l2 (final)		25.40
tilde-c-nmt	ensemble of 3 averaged models	**29.46**
tilde-c-nmt-comb	6 system combination	**29.36**
tilde-c-nmt-2bt	3 model average	27.99
tilde-nc-nmt (transf-u)	4 model average	**30.94**

Table 4: Automatic evaluation results

our unconstrained scenario systems (*tilde-nc-nmt*) ranked significantly higher than any other submission for both translation directions. Our best constrained systems were the second highest ranked systems among all constrained scenario systems, at the same time sharing the same cluster with the highest ranked systems.

6 Conclusion

The paper described the development process of the Tilde's NMT systems that were submitted for the WMT 2018 shared task on news translation. We compared Transformer models to MLSTM-based models and showed that the Transformer models outperform the older NMT architecture. We also showed that double back-translation may improve translation quality further than single back-translation. In terms of model ensembling and averaging, we showed that the best results in the constrained scenario were achieved by ensembling different run averaged models. In total, seven systems were submitted by Tilde for the English↔Estonian language pair.

	System	BLEU	DA	Cluster
English-Estonian				
	nict	25.16	62.1	2
(C)	tilde-c-nmt	23.54	61.6	2
	aalto	20.66	58.6	5
	tilde-nc-nmt	24.35	64.9	1
(U)	online-b	18.71	52.1	10
	neurotolge.ee	15.53	45.7	11
Estonian-English				
	nict	30.68	71.1	2
(C)	tilde-c-nmt	29.46	69.9	2
	uedin	29.38	69.2	2
	tilde-nc-nmt	30.94	73.3	1
(U)	online-b	25.81	67.1	2
	online-a	22.44	65.4	10

Table 5: Top three systems for the constrained (C) and unconstrained (U) scenarios according to the official results of the WMT 2018 shared task on news translation; ordered by the direct assessment (DA) standardized mean score

Acknowledgements

The research has been supported by the European Regional Development Fund within the research project "Neural Network Modelling for Inflected Natural Languages" No. 1.1.1.1/16/A/215.

References

Ahmet Aker, Monica Lestari Paramita, Mārcis Pinnis, and Robert Gaizauskas. 2014. Bilingual Dictionaries for All EU Languages. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)*, pages 2839–2845, Reykjavik, Iceland. European Language Resources Association (ELRA).

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. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

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. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical 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.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014. Jane: Open source machine translation system combination. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *ArXiv e-prints*.

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, 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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 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.

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 (WMT 2018), Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Mārcis Pinnis and Rihards Kalniņš. 2018. Developing a Neural Machine Translation Service for the 2017-2018 European Union Presidency. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 2018), vol. 2: MT Users*, pages 72–83, Boston, USA. Association for Machine Translation in the Americas.

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)*, volume 10415 LNAI, Prague, Czechia.

Mārcis Pinnis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics. 2017c. 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, Andrejs Vasiļjevs, Rihards Kalniņš, Roberts Rozis, Raivis Skadiņš, and Valters Šics. 2018. 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).

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

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.

Matīss Rikters, Chantal Amrhein, Maksym Del, and Mark Fishel. 2017. C-3MA: Tartu-Riga-Zurich Translation Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017a. The university of edinburgh's neural mt systems for wmt17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Others. 2017b. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68.

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)*, Berlin, Germany. Association for Computational Linguistics.

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.