eTranslation's Submissions to the WMT 2019 News Translation Task

Csaba Oravecz
IRIS Luxembourg
oravecz.csaba@gmail.com

Katina Bontcheva
Sogeti Luxembourg
katina.bontcheva@sogeti.lu

Adrien Lardilleux
C-Dev Luxembourg
adrien.lardilleux@c-dev.eu

László Tihanyi
IRIS Luxembourg
tihanyi1123@gmail.com

Andreas Eisele
DGT, European Commission
andreas.eisele@ec.europa.eu

Abstract

This paper describes the submissions of the eTranslation team to the WMT 2019 news translation shared task. The systems have been developed with the aim of identifying and following rather than establishing best practices, under the constraints imposed by a low resource training and decoding environment normally used for our production systems. Thus most of the findings and results are transferable to systems used in the eTranslation service. Evaluations suggest that this approach is able to produce decent models with good performance and speed without the overhead of using prohibitively deep and complex architectures.

1 Introduction

The European Commission's eTranslation[1] project, a building block of the Connecting Europe Facility (CEF), has been set up to help European and national public administrations exchange information across language barriers in the EU. It provides secure access to machine translation (both formatted documents and text snippets) between all 26 official languages of the EU and the EEA for translators and officials in EU and national authorities. In addition it enables multilinguality in all Digital Service Infrastructures of CEF.

CEF eTranslation builds on the previous machine translation service of the European Commission, MT@EC (Eisele, 2017), developed by the Directorate-General for Translation (DGT) since 2010. MT@EC translation engines were trained using the vast Euramis translation memories (Steinberger et al., 2014), comprising over 1 billion sentences in the 24 official EU languages, produced by the translators of the EU institutions over the past decades. While this large set of training data provides very good coverage of the type of language used in official EU documents, recent usage of the service is trending towards texts from other domains. The eTranslation team is working to widen the scope of the service and improve the coverage in more general types of texts. Given this background, the participation of eTranslation in this year's shared task on news translation is an early, but important step on a longer path towards a more generic MT service.

We participated in the task with 4 different language pairs: English→German, French→German, English→Lithuanian and Russian→English, in order to find best practices that guarantee the production of a solid system in a constrained resource environment.

2 Data Preparation

In this section we describe the data sets, the selection, and filtering methods that we applied to the provided parallel and monolingual data in order to increase the quality of trained models. We primarily focused on constrained submissions and made limited experiments with unconstrained resources, which we briefly describe later in Section 4.5.

2.1 Data Selection and Filtering

In most cases we used all of the provided original parallel data to build baseline models for back-

[1] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 320–326
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

Data set	En→De	Fr→De	En→Lt	Ru→En
Europarl v9	1.80M	1.72M	0.63M	–
Common Crawl	2.32M	0.62M	–	0.88M
News Commentary v14	0.32M	0.26M	–	0.29M
Rapid Corpus	1.47M	–	0.21M	–
Wiki Titles v1	1.25M	–	0.13M	1.00M
Yandex	–	–	–	1.00M
Total (unique):	7.16M (6.85M)	2.60M (2.59M)	0.97M (0.84M)	3.2M (2.1M)

Table 1: Number of segments in the filtered parallel data used for baseline models.

translation as well as for cross-entropy based filtering. The domain distribution of these data sets is not uniform across language pairs, which had some effect on the workflows we applied to specific language pairs. The basic procedure of data cleaning, however, was similar in all cases.

As a general clean-up, we performed the following steps on the parallel data:

- language identification with Python's `langid` module,
- segment deduplication with masked numerals,
- deletion of segments where source/target token ratio exceeds 1:3 (or 3:1),
- deletion of segments longer than 110 tokens,
- exclusion of segments without alphabetic characters.

The above steps reduced the data set by about 10%. However, we filtered out 65% of the Ru→En Wiki Titles corpus with an additional rule of having a minimum sum of 12 tokens in a segment pair. The number of segments in the base filtered data is shown in Table 1.

For the three language pairs[2] where we used monolingual data to build language models or create synthetic parallel text, we chose the recent target language News Crawl data sets, except for the 2018 German set, which contained a large number of segments with suspiciously scrambled characters in all words. Therefore, we discarded this version and made use of the 2016 and 2017 sets. In addition, for Fr→De we experimented with the 2014 and 2016 News Crawl as candidate data for the topic modeling based data selection (see Section 3.2). In the monolingual data used for backtranslation we performed some additional filtering; we set a threshold on the maximum length of a token (40) and the minimum ratio of letters to digits in a segment (4).

We applied dual conditional cross-entropy filtering (Junczys-Dowmunt, 2018a) to the provided ParaCrawl and CommonCrawl parallel datasets using the baseline translation models. This significantly reduced the size of these data sets without a major decrease in BLEU score for the high resource language pairs. For En→De the reduction in ParaCrawl was from 31M to 18M segments and in CommonCrawl from 2.3M to 1.4M segments with a drop of 0.2 BLEU points compared to using the full sets[3]. No additional cleaning was applied to the Fr→De and Ru→En Common Crawl since these already contained fewer than 1M segments. Experiments with the filtered (7.5M) and full (11M) ParaCrawl for Fr→De showed that the scores on the development test set were also almost identical. Therefore, we worked with this reduced data in the experiments to save time and resources. The parallel data for En→Lt was very small and we found that the unfiltered ParaCrawl was more beneficial than the filtered one. For Ru→En we used only the filtered ParaCrawl because we did not have time for more experiments.

Depending on data availability we opted for different ways of creating development and test data sets. For En→De we used the 2017 test set as validation set in the trainings and the 2018 test set as the test set to evaluate the trained models. For Fr→De we used the 2008–2014 test sets and

[2] Four minus Ru→En.

[3] This suggests that version 3 of the ParaCrawl is significantly less noisy than previous versions: we did not experience any improvement from filtering contrary to some of last year's experiments (Pham et al., 2018; Junczys-Dowmunt, 2018b).

randomly extracted 2000 segments for validation and 3000 segments for test, while the rest (about 13000) was kept for fine-tuning. For En→Lt we used a small random subset of the training data for validation and the provided development test for testing. For Ru→En we used the 2018 test set for testing and for validation we randomly extracted 3000 segments from the 2016 and 2017 newstests. The rest of the development data was used for fine-tuning.

2.2 Pre- and Postprocessing

The in-house translation workflow in the MT environment of eTranslation contains a fairly complex pre- and postprocessing pipeline, where standard steps (tokenization, normalization, placeholder replacement) are tailored to the Euramis data. It thus does not altogether fit the more heterogenous domain of WMT news data. This was confirmed in a few early baseline experiments on WMT 2018 parallel data where we simply used SentencePiece (Kudo, 2018), which allows raw text input/output within the the Marian toolkit (Junczys-Dowmunt et al., 2018). Since it proved to be superior to other (external) pre- and postprocessing workflows, we opted for this approach[4] in the 2019 experiments.

3 Trainings

Due to our low resource environment (no large-scale computing facilities), we did not have much room for experimenting with either a wide range of scenarios or much tuning of hyperparameters. Therefore, we decided to stick to simple setups and training procedures. In all experiments we used Marian, which is also the core of our standard NMT framework in the eTranslation service. All trainings were run as multi-GPU trainings on 4 NVIDIA P100 GPUs with 16GB RAM.

3.1 NMT Models

We trained only base transformer models (Vaswani et al., 2017) in all language pairs except for Fr→De and En→Lt, where we also tried experimenting with a big transformer.[5] We discarded the idea of building large ensembles of big transformers for high resource language pairs in the beginning due to the constrained environment. For most of the hyperparameters we used the default settings for the base transformer architecture in Marian[6] with dynamic batching and tying all embeddings. To save time and resources we stopped the trainings if sentence-wise normalized cross-entropy on the validation set did not improve in 5 consecutive validation steps. In the big transformer experiments, following recommended settings for Marian, we doubled the filter size and the number of heads, decreased the learning rate from 0.0003 to 0.0002 and halved the update value for `--lr-warmup` and `--lr-decay-inv-sqrt`.

Based on the results of previous experiments we set 30k joint SentencePiece vocabulary for En→De. We did not run additional trainings to test the effect of other vocabulary sizes, except for Ru→En, where we ran a baseline model experiment with separate 60k vocabularies. However, this resulted in a loss of 0.7 BLEU points on the 2018 test set.[7] Therefore, we kept the 30k joint setting through all language pairs.

3.2 Improving Baseline Models

In this section we describe the methods we experimented with to improve baseline models such as building an additional synthetic data set with back-translation (Sennrich et al., 2016), using the development data (where available) to fine-tune converged models with continued trainings and building ensembles out of a few variants of the best models originally trained from different seeds. We report the evaluation scores in Section 4.

3.2.1 Synthetic Data

Back-translation (Sennrich et al., 2016) has become a widely used data augmenting technique in NMT but at the same time significantly extends the search space for best settings as far as the amount of data, ratio of bitext to back-translation data or methods to generate the synthetic source are concerned (Edunov et al., 2018).

In the En→De system we experimented with adding 10M and 20M back-translated segments from the 2017 News Crawl to the available bitext. The latter setting yielded no improvement, in ef-

[4] We used default settings for Marian's built-in SentencePiece: unigram model, built-in normalization and no subword regularization.

[5] However, the difference between the base and big tranformer models for Fr→De and En→Lt was not significant. We decided to submit the big models in the hope of their better performance on the shared task test set.

[6] See eg. `https://github.com/marian-nmt/marian-examples/tree/master/transformer`.

[7] Confirmed post-submission with a loss of 1.9 BLEU points on the 2019 test set.

fect it was slightly worse so for the final systems the 10M data set was used. We had no time and resources for more fine-grained experiments to find the optimal setups with back-translation data.

For Fr→De we tuned our models towards the topic defined in the task by making use of guided topic modeling[8]. We manually created a seed word list with around 100 tokens from a few German news articles on elections, then we classified the documents in the 2014 and 2016 German News Crawl data sets into different topics.[9] We finally selected about 170k doc units from News Crawl 2014 and 186k doc units from News Crawl 2016 as candidate data for back-translation. We also experimented with back-translation of 2.5M randomly selected segments from News Crawl 2017. This synthetic data brought some improvement but not as much as the synthetic data obtained from topic modeling.

For En→Lt we back-translated all of the provided monolingual data with the exception of Common Crawl. We filtered Common Crawl using a language model built on the only in-domain resource for this language pair, 2018 News Crawl. We took the top 500k segments and back-translated them but this did not result in any improvement (we used, however, a transformer type language model built on 2018 News Crawl for later models (cf. Section 3.2.3)).

3.2.2 Fine-tuning with In-domain Data

For language pairs where a substantial amount of test data from previous years' tasks is available a possible direction to improve performance is to continue training with this data as domain adaptation (Luong and Manning, 2015). For En→De we used the 2008–2017 development sets (30k segments) in the experiments and for the final submission we extended it with the 2018 test set. For Ru→En we used a set of about 18k segments from the news test sets from 2012 onwards, with the exception of the data used for testing and validation.

In the Fr→De system we used a set of about 13k segments (cf. Section 2.1). It yielded improvements on our test set, which was selected randomly rather than through topic modeling. Since we tuned the system this way towards the more general news domain it is not surprising that for the 2019 test set this fine tuning proved to be harmful. Unfortunately, we submitted the fine-tuned model, which, although it did not alter our position in the rankings, still led to a loss of 0.8 BLEU points (cf. Table 3 in Section 4.2).

3.2.3 Ensembles

For the final En→De submission we created a 3 model ensemble trained with the same (best) configuration but with different seeds. We also built an ensemble with a transformer type language model from the 2016 and 2017 German News Crawl (117M segments) which we trained for 2 epochs. We set the weight of the language model to 0.1 and the weight of the translation models to 1.0 to get the largest improvement.[10]

For the Fr→De and En→Lt final submissions, we also created ensembles from the best single models trained from different seeds but here we only had time to experiment with 2 models. For En→Lt we added a transformer type language model from filtered 2018 news (375k) to the ensemble. Similarly to En→De, the translation models had a weight of 1.0, while the language model had a weight of 0.1.

3.2.4 Ineffective Methods

We make a brief mention of the methods that we tried but did not seem to work. In particular, for En→De oversampling the original parallel data did not yield any improvement so we stopped the experiments in this direction. Since for Ru→En the addition of the UN corpus did not increase model quality, we left it out from the training data.[11] Another technique that seemed promising but did not give any improvement was incremental iterative back-translation (Hoang et al., 2018; Marie et al., 2018). For En→Lt, where the available data set was in general much smaller, we had time to experiment with this technique but we did not manage to generate better models.

4 Results

We submitted one model for each of the four language pairs. In this section we provide evaluation scores for models at important stages in the experiments which reflect how the models got better as

[8] https://github.com/vi3k6i5/guidedlda

[9] We tokenized the text and used stopword list but no lemmatization in creating the document-term matrices.

[10] It might be worth noting that the ensemble of a single translation model and the language model did not give any improvement; a small increase in the final score could only be obtained by adding the language model to the 3 member ensemble.

[11] Some WMT 2018 participants had similar experience in the En→Ru direction (Deng et al., 2018).

we tried various methods for improvement. All results are reported in detokenized BLEU.[12]

4.1 English→German

System	Parallel data	2018	2019
M1 Baseline	6.8M	41.3	38.1
M2 M1+PC	24M	44.6	39.9
M3 M2+BT	34M	45.4	38.7
M4 M3 ens.	34M	46.0	40.1
M5 M4+LM	34M	46.3	40.3
M6 M5+FT	34+0.03M	**47.8**	**42.4**

Table 2: Results for En→De models. The 2019 results are post-submission.

Table 2 summarizes the scores for the En→De models. Model 1 as our baseline used only the original parallel data (Table 1). In Model 2 we extended this data with filtered ParaCrawl (PC) v3 data, which led to a substantial improvement (although less so on the 2019 test set). For Model 3 we added the synthetic data (BT), which seemed to improve the quality on the 2018 test set but to our great surprise resulted in a performance drop on the 2019 test set. This might suggest that the synthetic data already introduces some unwanted noise into the model that could have a detrimental effect depending on the input to be translated. Model 4 is an ensemble of three Model 3 setups and this proved to be a very efficient choice with respect to the 2019 test set. Some small additional improvement could be gained by adding the language model (LM) to the ensemble (Model 5) but the largest positive effect came from the fine tuning (FT) as seen in Model 6.

4.2 French→German

Table 3 gives the scores for the Fr→De models. The 2008-14D column contains the scores on our development test set (cf. Section 2.1). The baseline Model 1 is built from the original parallel data (Table 1). In Model 2 we added a small amount of back-translated data, which was generated from the monolingual Europarl and News Commentary. From this data set we filtered out the segments that overlap with the original parallel data. This step led to a moderate improvement. For Model 3 we added filtered ParaCrawl v3 data, again with a moderate improvement. In Model 4 we included the topic selected synthetic data, which improved the quality minimally on the development set but significantly on the 2019 test set. In Model 5 we fine-tuned Model 4, which gave yet again a moderate improvement on the development set but resulted in a decrease on the 2019 test set (cf. Section 3.2.2). At this stage, we decided to test big transformers from Model 4. We only had time to train 2 models and even they could not reach convergence in time. Model 6 is an ensemble of the 2 big transformers, each with a weight of 1.0, while for Model 7 we ensembled the fine-tuned models of Model 6. Unsurprisingly, Model 7 was better than Model 6 on the development set but worse on the 2019 test data (cf. Section 3.2.2). For this language pair, the most beneficial step was the addition of topic-selected back-translated data.

System	Parallel data	2008-14D	2019
M1 Baseline	2.6M	20.8	26.1
M2 M1+BT1	3.2M	21.4	27.8
M3 M2+PC	6.9M	22.4	29.4
M4 M3+BT2	11.6	22.8	33.1
M5 M4+FT	11.6M+13k	23.8	32.4
M6 M4 ens.	11.6M	22.7	**33.5**
M7 M5 ens.	11.6+13k	**24.3**	32.7

Table 3: Results for Fr→De models. The 2019 results are post-submission.

4.3 English→Lithuanian

System	Parallel data	2019D	2019
M1 Baseline	0.84M	15.5	11.4
M2 M1+PC	2.2M	19.4	12.5
M3 M2+BT	4.7M	25.7	16.6
M4 M3+OS	5.9M	25.8	15.9
M5 M4+LM	5.9M	26.1	16.0
M6 M5 ens.	5.9M	**27.0**	**17.1**

Table 4: Results for En→Lt models. The 2019 results are post-submission.

Table 4 presents the scores for En→Lt. The 2019D column is for the scores on the provided development set (cf. Section 2.1). Model 1 is the baseline with the original parallel data (Table 1). In Model 2 we added the full ParaCrawl

[12] sacreBLEU signatures: `BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+tok.13a+version.1.3.0`

v3 data, which led to a substantial improvement on the 2019 development set but just a moderate one on the 2019 test set. In Model 3 we further added the synthetic data (back-translation of all monolingual data except Common Crawl). This resulted in a big boost in the quality on both test sets. For Model 4 we oversampled (OS) 2 times the Rapid corpus from the parallel data and the domain-relevant back-translated data (2018 News Crawl). Model 5 is a $(1, 0.1)$ ensemble of Model 4 with a transformer-type language model, with a minimal improvement on the 2018 development set but a drop of 0.6-0.7 BLEU points on the 2019 test set. Since this was unknown in the development stage, we decided to build big transformer models on the same training data as Model 4. Model 6 is an ensemble of these 2 big transformers and the language model. The improvement on the 2019 test set was significant.

4.4 Russian→English

System	Parallel data	2018	2019
M1 Baseline	2.1M	27.3	32.4
M2 M1+PC	5.9M	29.5	35.9
M3 M2+FT	5.9M+17.8k	**32.9**	**37.4**

Table 5: Results for Ru→En models. The 2019 results are post-submission.

We made fewer experiments with the Ru→En system. The scores in Table 5 give the outcome of the evaluation of three simple single transformer models: (i) Model 1 built on the original parallel data (excluding the UN corpus); (ii) Model 2 with filtered ParaCrawl added; (iii) Model 3, which is fine-tuned on domain-specific data. This shows that it is possible to produce reasonable models in very constrained conditions.

4.5 Experiments with Unconstrained Models

We ran a few experiments with unconstrained models making use of the Euramis (Steinberger et al., 2014) data set. This data contains millions of segments for 3 of the 4 language pairs we worked with and offers itself as a natural resource to build unconstrained models from. At the same time it is in general quite distant from the news domain. Thus for the high resource language pairs (En→De, Fr→De) we first tried to use only those subsets which might be closer to the shared task domain. We extracted additional training data using language models built from monolingual news corpora as reference in-domain text with the XenC toolkit (Rousseau, 2013). For Fr→De we built the language model from the topic modeling based selection and also experimented with extracting Euramis data using the same guided LDA process as described in Section 3.2.1. We re-ran the trainings of the best constrained models by adding 2M and later 3M Euramis segments to the training data but as we cannot report on any improvement, we stopped this line of experiments and did not submit the unconstrained systems.

For En→Lt, we trained 3 non-constrained models by adding to the best constrained system (i) all our Euramis data, (ii) 1M and (iii) 2M segment subsets selected as described above. This resulted in a very small improvement of less than 0.5 BLEU points for the models with selected Euramis subsets, while the model with the full Euramis data was almost 2 BLEU points worse. We thus decided not to submit any of the 3 systems.

5 Conclusion

For the first participation in WMT 2019, the eTranslation team submitted four systems to the news translation shared task. We experimented with different settings for each task but the development of all systems shared the common goal of maximizing efficiency in a relatively low-resource production environment. For this reason, our systems relied on simple architectures, and we focused instead on finding the most appropriate combination of standard techniques and tools, which can thus directly be ported to production systems. In particular, we could confirm that a careful selection of the training data, back-translation and fine-tuning were generally the most rewarding techniques, allowing all our systems to perform decently and to end up in the first half of the rankings, despite the limitations imposed by our low resource environment.

References

Yongchao Deng, Shanbo Cheng, Jun Lu, Kai Song, Jingang Wang, Shenglan Wu, Liang Yao, Guchun Zhang, Haibo Zhang, Pei Zhang, Changfeng Zhu, and Boxing Chen. 2018. Alibaba's neural machine translation systems for WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 372–380, Belgium, Brussels. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Andreas Eisele. 2017. Machine translation at the European Commission. In Jörg Porsiel, editor, *Machine Translation: What Language Professionals Need to Know*, pages 209–220. BDÜ Fachverlag, Berlin.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24. Association for Computational Linguistics.

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

Marcin Junczys-Dowmunt. 2018b. Microsoft's submission to the WMT2018 news translation task: How I learned to stop worrying and love the data. In *Proceedings of the Third Conference on Machine Translation*, pages 429–434, 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. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 76–79.

Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2018. NICT's neural and statistical machine translation systems for the WMT18 news translation task. In *Proceedings of the Third Conference on Machine Translation*, pages 453–459, Belgium, Brussels. Association for Computational Linguistics.

Ngoc-Quan Pham, Jan Niehues, and Alexander Waibel. 2018. The Karlsruhe Institute of Technology systems for the news translation task in WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, pages 471–476, Belgium, Brussels. Association for Computational Linguistics.

Anthony Rousseau. 2013. XenC: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100:73–82.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 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 5998–6008. Curran Associates, Inc.