eTranslation's Submissions to the WMT22 General Machine Translation Task

Csaba Oravecz[⋆] **Katina Bontcheva**[†] **David Kolovratník**[†]
Bogomil Kovachev[⋆] **Christopher Scott**[⋆]
DG Translation – DG CNECT, European Commission
[⋆]firstname.lastname@ec.europa.eu
[†]firstname.lastname@ext.ec.europa.eu

Abstract

The paper describes the 3 NMT models submitted by the eTranslation team to the WMT22 general machine translation shared task. In the WMT news task last year, multilingual systems with deep and complex architectures utilizing immense amounts of data and resources were dominant. This year with the task extended to cover less domain specific text we expected even more dominance of such systems. In the hope to produce competitive (constrained) systems despite our limited resources, this time we selected only medium resource language pairs, which are serviced in the European Commission's eTranslation system. We took the approach of exploring less resource intensive strategies focusing on data selection and data filtering to improve the performance of baseline systems. With our submitted systems our approach scored competitively according to the automatic rankings in the constrained category, except for the En→Ru model where our submission was only a baseline reference model developed as a by-product of the multilingual setup we built focusing primarily on the En→Uk language pair.

1 Introduction

The eTranslation team is responsible for the development of machine translation systems providing the translation services of the European Commission's eTranslation project[1]. This is a building block of the Connecting Europe Facility (CEF), with the aim of supporting European and national public administrations' information exchange across language barriers in the EU. The project is described in more details in Oravecz et al. (2019).

During the previous years the team's participation in the WMT shared tasks allowed us to explore state-of-the-art methods to develop high quality machine translation systems. However, due to strict resource constraints, these systems do not normally carry over to production environments and there has been a continuous search for the right balance between the use of resources in production environments and the best performing but more complex architectures.

With the news translation shared task extended to being a general MT task the need for more robustness, coverage and consequently more complexity and resources has further increased. We expect a strong competition in these areas, where teams with modest resources might have some inherent disadvantages. Therefore, in this year's experiments we did not consider high resource language pairs (specifically English → German, our constant submission in previous years) and opted for the medium resource French → German and English → Ukrainian language directions. The latter system originated from a multilingual setup including Russian data, so we built and submitted a baseline English → Russian model as well.

2 Data Preparation

In this section we briefly describe the base data sets, the general selection and filtering methods we applied to prepare these initial data sets used to train the first baseline models. Further data selection and augmentation methods to improve the quality of baseline models are described in Section 3.1. We only used the provided parallel and monolingual data, so our submissions all fall into the constrained category.

2.1 Base Data Selection and Filtering

As a general clean-up, we performed the following filtering steps on the parallel data[2]:

[1] https://language-tools.ec.europa.eu

[2] In some subcorpora, only a subset (not necessarily the same) of these steps was applied, depending on the data set. No filtering was used for the dev sets.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 346 - 351
December 7-8, 2022 ©2022 Association for Computational Linguistics

Data set	Fr→De	En→Uk, Ru	En→Uk	En→Ru
Europarl v10	1.79M	–	–	–
Common Crawl	0.42M	0.78M	–	0.78M
News Commentary v16	0.29M	0.34M	–	0.34M
Tilde Model Corpus	4.24M	9.00k	1.00k	8.00k
Dev sets	0.03M	–	–	–
Wiki Titles v3	0.99M	0.70M		0.70M
ParaCrawl	5.64M	12.9M	7.60M	5.30M
OPUS	–	22.9M	22.9M	–
WikiMatrix	1.99M	5.28M	1.50M	3.78M
Yandex	–	1.00M	–	1.00M
UN Parallel	–	9.19M	–	9.19M
Total:	15.39M	53.1M	32.0M	21.1M

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

- language identification with FastText[3] (Joulin et al., 2016),
- segment deduplication,
- deletion of segments where source/target token ratio exceeds 1:3 (or 3:1),
- deletion of segments longer than 100-150 tokens (depending on language pair),
- exclusion of segments where the ratio between the number of characters and the number of words was below 1.5 or above 40,
- exclusion of segments without a minimum number of alphabetic characters (2),
- exclusion of segments with tokens longer than 40 characters,
- exclusion of segments where the length difference between source and target in the number of tokens was higher than 8,
- removal of segments where source side contained specific noise patterns (in Fr→De ParaCrawl).

These filtering steps led to an average reduction of about 15-20% of the training data with the number of segments as shown in Table 1. For Fr→De, after some manual inspection of the raw WikiMatrix and ParaCrawl data, we decided to experiment with some further clean-up on these data sets, using dual conditional cross-entropy filtering (Junczys-Dowmunt, 2018), where we built the scoring models from a subset of filtered parallel data (7.6M segments) by excluding ParaCrawl and WikiMatrix. We then built models by deleting the worst scoring 5 and 10 % of the two data sets but none of these models was better then the baseline system, so we did not use this filtering in the submission setups. In En→Uk, we experimented with language model based filtering, where we built the language model from the Leipzig corpora and fine tuned the baseline model on the filtered data set, however, it gave no improvement, so this step was not used in the submission systems either.

2.1.1 Monolingual data

In the Fr→De models, where we used back-translation (Sennrich et al., 2016) to improve baseline performance we utilized monolingual data from the various corpora provided. The data was filtered with the same rules (where applicable) as the parallel data (see Section 2.1). Table 2 provides a summary. For the other systems, we didn't use back-translated data in the submissions[4], only the original parallel data sets.

2.1.2 Development and test data

For Fr→De, since the task had been extended from news translation to general MT, where test data was expected from "news, e-commerce, social, and conversational" text, we opted to use a custom built

[3] https://fasttext.cc/docs/en/language-identification.html

[4] See Section 3.1 for experiments with monolingual Ukrainian news. The other monolingual Ukrainian data sets that could have been used for back-translation came too late for us to be able to reschedule the trainings.

Data set	Fr→De
Europarl v10	2.08M
Leipzig mixed	0.99M
Leipzig web	0.99M
News Commentary v16	0.43M
News Crawl 2021	25.0M
Total:	29.29M

Table 2: Number of segments in the filtered monolingual data used for back-translation.

test set for development rather than some previous dev set from the news domain. We extracted a 10k random subset from the filtered original parallel data and manually selected 2k segments for test and validation each. In the manual selection we tried our best to keep segments most representative of the expected domains. These segments were then obviously removed from the training data.

For En→Uk, the validation data was extended with 2k segment pairs randomly extracted from the filtered original parallel data. In addition to the Flores test set, we used 2 development test sets: 10k segment pairs extracted at random from OPUS, and 5k segment pairs extracted from ParaCrawl.

For En→Ru, we extended the validation data again with 2k segment pairs extracted at random from the 2012–2020 dev sets. Beside the Flores test set, we used 2 additional test sets: a 5k random extraction from the parallel data and the provided 2021 news test set. In the latter two language pairs we did not apply manual selection, we considered the test sets already representative enough for the task.

2.2 Pre- and Postprocessing

As in our previous years' systems, we applied the simplest possible workflow without the standard pre- and postprocessing steps of truecasing, or (de)tokenization, and simply used SentencePiece (Kudo, 2018), which allows raw text input/output within the Marian toolkit (Junczys-Dowmunt et al., 2018)[5] in the experiments. In the submission hypotheses, some simple normalization steps were applied in post-processing, similarly to previous years.

3 Trainings

In all experiments we used Marian, as the core tool of our standard NMT framework in the eTranslation service. Trainings were run as multi-GPU setups on 4 NVIDIA V100 GPUs with 16GB RAM, typically for about 30 epochs. In general, except for the first baseline setups, we built only big transformer models, this year even for back-translation, in the hope of getting better quality output for the higher resource consumption. The development scenario was straightforward without much room for experimenting with different parameters or setups due to limited resource availability: for Fr→De, a single set of 4 member ensembles from big transformers, while in En→Uk and En→Ru, a multilingual model at the first stage, fine tuned on the specific languages at the second stage, with 4 (Uk) and 3 (Ru) models in an ensemble as our submission systems for these two language pairs. The parameter settings did not change from last year's setup: for most of the hyperparameters we used the default settings in the baseline models for the base transformer architecture in Marian[6] with dynamic batching and tying all embeddings. In Fr→De, trainings were stopped if sentence-wise normalized cross-entropy on the validation set did not improve in 5 consecutive validation steps. The multilingual systems were stopped after about 40 epochs, and then fine tuned for each target direction until they were stopped to meet the submission deadline.

In the big transformer setups, we also followed standard settings for Marian, i.e. 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`.

Following common ranges of subword vocabulary sizes, we set a 32k joint SentencePiece vocabulary for all language pairs. SentencePiece models were trained from 10M random segments.

3.1 Synthetic Data

In Fr→De, we back-translated the monolingual data described in Section 2.1.1 with a single big transformer trained from all available original parallel data. The resulting synthetic data set was filtered (where applicable) with the same techniques as the original parallel data. To train the submission

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

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

		Test sets					
		Dev			2022		
System	Data	COMET	ChrF	BLEU	COMET	ChrF	BLEU
M1: Bilingual baseline	32.0M	69.9	63.2	40.3	47.4	52.9	24.4
M2: Multilingual En→{Uk,Ru}	53.1M	68.4	62.3	39.2	46.7	52.5	24.2
M3: M2 fine-tuned on En→Uk	53.1M/32.0M	71.2	63.7	40.9	50.1	53.3	24.5
M4: M2^{bigTr}	53.1M	73.0	64.2	41.5	53.3	54.3	25.6
M5: M4 fine-tuned on En→Uk^{bigTr}	53.1M/32.0M	74.2	65.0	42.7	52.5	54.4	25.8
M6: 4 x M5 ens.$^{bigTr}_{subm}$	53.1M/32.0M	75.0	65.3	43.2	**54.5**	**54.8**	**26.2**

Table 3: Results for En→Uk models. The *Dev* column displays the global scores for all dev sets concatenated.

ready systems we upsampled the (filtered) baseline original parallel (OP) data set to a 1:1 ratio with the BT data (Ng et al., 2019; Junczys-Dowmunt, 2019). This setup was a one shot configuration, we lacked the resources to experiment with other OP-BT combinations. As in previous years, we used tagged back-translation (Caswell et al., 2019) in our workflows.

In En→Uk, back-translation of a 2.4 M subset of monolingual news data with a reverse engine trained from original parallel data did not yield any improvement over the baseline so it was not used in the submission systems.

3.2 Continued Trainings

For Fr→De, in the first phase of the trainings we used all available OP data together with the back-translated synthetic data set. As a second phase after model convergence, we continued the training for 3 additional epochs[7] only on the OP data set.

In the multilingual setup, the first phase of the trainings utilized all available OP data for En→Ru and En→Uk[8]. These trainings were stopped after about 40 epochs and continued only on the respective target data. In both phases the source language data was prefixed with the target language code. All continued trainings were stopped before the submission deadline.

4 Results

We submitted a constrained system for each of the 3 language pairs. We provide COMET (Rei et al., 2020) (with the default model `wmt20-comet-da`), ChrF (Popović, 2017) and BLEU (Papineni et al., 2002) evaluation scores for models at important stages in the development, which reflect how the performance of the models changed as we experimented with the various configurations.[9]

4.1 English→Ukrainian

Table 3 gives a summary of the of the En→Uk experiments. The baseline model (M1) was trained on the filtered original parallel (OP) data using the base transformer architecture. We did not primarily go for a system with synthetic data since the usable monolingual Uk data was small in size (2.6M after filtering) and we didn't expect substantial improvement. Instead, we decided to experiment with multilingual systems. The next model (M2) was a multilingual En→{Uk,Ru} system trained only on filtered OP data (En→Uk, En→Ru), again as a base transformer. The target language was indicated in a token that was prefixed to the source language segments. The slight drop of the scores compared to M1 is not unexpected in multilingual NMT systems when using the same architecture as the bilingual model (Wang et al., 2020). In the next step we used the model of M2 that scored best on the En→Uk development test sets and fine-tuned on En→Uk data until convergence (early-stopping set to 20 stalls). This fine-tuned model was better than the bilingual baseline (M1) and the multilingual M2. The next step (M4) was to train M2 with big transformer architecture. This model was significantly better than all 3 previous models. M5 was an M4 model fine-tuned on En→Uk data, while M6 (our submission model) was a 4 member ensemble built from M5 models. Both M5 and M6 yielded some

[7] We experimented with different number of epochs, until we saw a steady improvement on the test set.

[8] Without EU-Acts, which came too late.

[9] sacreBLEU signatures:
`chrF2|nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0`
`BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0`

		Test sets					
		Dev			2022		
System	Data	COMET	ChrF	BLEU	COMET	ChrF	BLEU
M1: Bilingual baseline	21.1M	51.2	57.2	31.8	48.3	53.4	27.0
M2: Multilingual En→{Uk,Ru}	53.1M	50.3	56.9	31.1	47.3	53.1	26.7
M3: M2^{bigTr}	53.1M	57.8	59.5	34.1	56.2	55.4	29.2
M4: M3 fine-tuned on En→Ru^{bigTr}	53.1M/21.1M	59.6	59.9	34.8	56.1	55.3	29.1
M5: 3 x M4 ens.$^{bigTr}_{subm}$	53.1M/21.1M	60.3	60.3	35.3	**57.9**	**55.8**	**29.8**

Table 4: Results for En→Ru models. The *Dev* column displays the global scores for all dev sets concatenated.

		Test sets					
		Dev			2022		
System	Data	COMET	ChrF	BLEU	COMET	ChrF	BLEU
M1: Baseline	15.4M	64.6	62.1	32.9	47.2	65.2	41.5
M2: M1+BT^{bigTr}	59.4M	65.1	62.3	33.0	53.1	67.1	44.5
M3: M2 cont.^{bigTr}	59.4M	65.5	62.4	33.1	53.4	67.3	44.9
M4: 4 x M3 ens.$^{bigTr}_{subm}$	59.4M	66.7	62.8	34.0	**55.4**	**68.4**	**46.5**

Table 5: Results for Fr→De models.

improvement in the automatic metrics.

4.2 English→Russian

The main stages of the model development for the En→Ru language pair are presented in Table 4. As we described before, the En→Ru system was not intended to be a competitive submission, and this is reflected in the evaluation scores, which are below the scores of other submissions. The baseline model (M1) was trained on the filtered OP data as a base transformer. The next two models (M2 and M3) are common with En→Uk (M2 and M4) – a multilingual En→{Uk,Ru} systems trained only on filtered OP data as base/big transformers (cf. Section 4.1 above). M4 is the M3 model fine-tuned on En→Ru OP data, while M5 (our submission model) is a 3 member ensemble built from M4 models. The score improvements are similar to En→Uk.

4.3 French→German

Table 5 summarizes the results of the Fr→De experiments. The first baseline model (M1) was trained only on the (filtered) original parallel (OP) data with the base transformer architecture. The next model (M2) switched to the big transformer setup and used the back-translated (BT) data with the OP data upsampled (see Section 3.1). Despite the significant increase of the training data size, the effect on the scores on our development set was moderate, however, on the 2022 test set the increase was substantial. This might suggest that the back-translated data gave better support than the OP data to the 2022 test set as a general test set but was much less effective for our development set (which was perhaps still too restricted to the news domain). In the 3rd model (M3), we continued the training only with the OP data as described in Section 3.2, with a slight increase in the metrics. Our submission model (M4) was a 4 member ensemble built from M3 models, where the 4th model was weighted 10% more than the rest. This configuration yielded the most promising result with a significant increase in the scores suggesting that ensembling might be an efficient strategy for general MT models.[10] Model 4 ended up as the best submission of the constrained category, according to all automatic metrics.

5 Conclusion

We described the submissions of the eTranslation team to the WMT22 general MT shared task on 3 language pairs: French-German, English–

[10] In previous years, ensembling was less efficient in our submitted news specific models.

Ukrainian and English–Russian, the last submission being only a baseline setup for reference, built only as a by-product of the En→Uk system. We selected medium resource language pairs and tried to focus on data selection, filtering and evaluation with custom test sets to be able to produce strong constrained systems even with limited resources. In our two competitive systems, first automatic results seemed to justify this approach.

References

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

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

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. 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.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Csaba Oravecz, Katina Bontcheva, Adrien Lardilleux, László Tihanyi, and Andreas Eisele. 2019. eTranslation's submissions to the WMT 2019 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 320–326, Florence, Italy. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.