# eTranslation's Submissions to the COVID19-MLIA Translation Task

Csaba Oravecz† Katina Bontcheva† David Kolovratník† Bogomil Kovachev⋆
Vilmantas Liubinas⋆ Christopher Scott⋆ Francois Thunus⋆ Andreas Eisele⋆

DG Translation – DG CNECT, European Commission
†firstname.lastname@ext.ec.europa.eu
⋆firstname.lastname@ec.europa.eu

**Abstract.** The report describes the 6 NMT models submitted by the eTranslation team to the COVID19-MLIA translation shared task in Round 2. We developed systems in language pairs that are actively used in the European Commission's eTranslation service. We focused primarily on data filtering and applied standard techniques of fine tuning and ensembling to improve our models. The submitted systems scored competitively in all language pairs, and several of our models were the best according to the automatic evaluation.

## 1 Introduction

The eTranslation team submitted constrained and unconstrained systems for 6 language pairs. We used standard best practices to develop the models focusing on cleaning up the provided training data and finding optimal architectures for training. For the unconstrained models we generally tried to use all available health related data that we had access to, which in some language pairs led to a significant increase in model quality while in other cases the difference in performance was not substantial. The systems submitted and the experiments during the development are described in detail for each language pair in the following sections.

## 2 English→German

### 2.1 Data

For the second round of submissions the provided parallel data increased considerably. At the same time significant sections of the data set could hardly be considered clean parallel data, only comparable. Therefore we applied some basic rule based filtering as well as some simple heuristics to exclude noisy segments. As a general clean-up, we performed the following steps on the raw data set:

– language identification with FastText[1] [2],

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

– deletion of segments where source/target token ratio exceeds 1:3 (or 3:1),
– deletion of segments longer than 110 tokens,
– 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 (4) of alphabetic characters.

These filtering steps led to a 3% reduction of the data set (from 2,462,556 segments to 2,392,900).

Using some more elaborate approach based on powerful resources to remove segments where the target side was far from being the translation of the source was not a viable option for the constrained category, therefore we applied only a basic heuristic: if there was a mismatch in the numeric tokens[2] between source and target we removed the segment from the data set. This resulted in a further 6% reduction (from 2,392,900 to 2,249,452 segments). We did not perform a methodical evaluation on the performance of this filter but on a small scale manual evaluation its precision was very high, segments marked for removal were indeed non-parallel. Recall however was probably much lower, many noisy segments could still remain in the data set. However, we did not perform more filtering and used this set of 2,249,452 segments to build our models.

To test our models we used several different test sets from the data available, including a post submission in house test set created from health related Euramis segments (see below) and a 2k test set extracted from Round 2 validation data. Results were not homogeneous across models and test sets, some models performed better on one set but worse on another. Models selected for submission performed strong on the development test set so we mainly report results for these models.

### 2.2 Training

Similarly to Round 1, for the constrained experiments, we first built a base transformer model from the filtered training data. The training process was also similar, we split up the validation set into two halves and used one part to stop the trainings if sentence-wise normalized cross-entropy on the validation set did not improve in 5 consecutive validation steps, and reserved the other part as a test set. We did not apply any standard pre- and postprocessing steps of truecasing, or (de)tokenization; we simply used SentencePiece [4], which allows raw text input/output within the Marian toolkit [3][3] For most of the hyperparameters we used the default settings for the base transformer architecture in Marian[4] with dynamic batching and tying all embeddings. We experimented with big

[2] When checking mismatch, we removed all punctuations (eg. decimal comma or dot, slash or hyphen etc.) from the numerals.

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

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

transformer architectures, here we also followed 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`. We experimented with (joint) vocabulary sizes of 12k and 32k, the latter resulting in somewhat better scores. The submissions models used the latter setting. As a last step in the training, our best constrained model benefited from a 5 epoch fine-tuning on the Round 1 and 2 validation and Round 1 test sets.

The trainings were run as multi-GPU trainings on 2 (base transformer) or 4 (big transformer) NVIDIA V100 GPUs with 16GB RAM, for around 30 epochs.

### 2.3 Unconstrained models

Our first unconstrained submission is a single big transformer model trained from the filtered training set extended with the TAUS Corona Crisis Corpora[5] (610k segments), the OPUS EMEA Corpus[6] (760k segments), and a health related subset of the Euramis data set [5] (1.1M segments).

The second unconstrained submission as in Round 1 is based on and uses the strongest model that the eTranslation team submitted to the WMT 2021 News Task. This model is a 4 member big transformer ensemble, trained as a constrained submission for WMT on more than 400M original parallel and back-translated segments, with fine tuning on the development sets. Looking at the result of Round 2 we can confirm that what we hypothesized in Round 1 still holds: this model, although primarily oriented towards the news domain, is a powerful general MT system, and it significantly outperforms all other systems in the current task already in zero shot mode. In the submission model, each of the 4 big transformer models is fine tuned on the filtered training data for 3 epochs and ensembled together with equal weights. This model is marginally better than the zero shot variant. This and the large difference between the constrained and unconstrained models seem to suggest that the big WMT model already gives good support for the test set in the current task, and, possibly, the test set is still closer to the news domain in general than to the range of the (in-domain) training data (which itself might still be heterogeneous and as such too small to compete with the large general news model).

### 2.4 Results

In Table 1 we present the BLEU scores[7] for the main models we experimented with in the development of En→De systems. Models marked by an asterisk are submission models. We do not calculate the scores where the training data already contains the test segments of the particular test set. Results reported with

[5] `https://md.taus.net/corona`

[6] `https://opus.nlpl.eu/EMEA.php`

[7] sacreBLEU signatures: `BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+tok.13a+version.1.5.1`

| System | Data | Test sets | | |
|---|---|---|---|---|
| | | Euramis (2k) | R2-V (2k) | R2 off. (4k) |
| Round 1 (c) | 926k | 32.7 | 29.2 | 27.7 |
| Round 2 raw (c) | 2.46M | 35.3 | 39.9 | 39.7 |
| Round 2 filt. (c) | 2.25M | 35.9 | 40.3 | 39.7 |
| R 2 filt. big Tr. (c) | 2.25M | 35.3 | 39.0 | 38.3 |
| Big tr. ensemble* (c) | 2.25M | 37.3 | 41.7 | 40.9 |
| Big tr. ens. fine t.* (c) | 2.25M+6.5k | 37.3 | – | **41.1** |
| Single big Tr.* (u) | 3.89M | – | 41.3 | 41.2 |
| WMT21 (u) | 430M | 38.7 | 45.9 | 46.2 |
| WMT21 fine t.* (u) | 430+2.25M | 39.9 | 47.4 | **47.1** |

**Table 1.** Results for En→De models. R2-V is the development test set extracted from Round 2 validation data. (c): constrained, (u): unconstrained model. Submissions are marked with an asterisk.

the Round 2 test set (last column) for the submitted models are slightly higher than on the evaluation portal. The reason for this is that we submitted versions of the models' hypotheses with a postprocessing step normalizing German punctuation but as it turned out the reference set did not follow German typography, so to calculate the scores in Table 1 we rolled back to the raw model outputs. According to these results we achieved the highest scores in both the constrained and unconstrained categories.

The different test sets do not seem to give a clear, unanimous indication of the best workflows and model architectures, the various setups behave slightly differently on one or the other test set. From the development test sets it seems that data filtering was a useful step while the Round 2 test set gives identical scores for this setting. The switch from the base to the big transformer architectures was not, according to the scores, justified, although the ensemble clearly outperformed all other (constrained) models. It has been shown that in low resource scenarios moderate architectures perform better [1], and it seems even in medium resource settings the base transformer architecture remains competitive. Fine tuning with the development sets in the constrained settings yielded a small increase but in the unconstrained models the same technique using the full in-domain training set led to a more significant improvement.

## 3 English→Swedish

### 3.1 Data

Training data was made up of segments from both Round 1 and Round 2, while validation data only came from what was provided for Round 2. After sampling the data manually for quality we performed two clean up tasks to filter out certain problematic segments. The nature of the crawled data meant that in many cases, numbers and locations did not match between segments and these

were removed where possible using scripts. This reduced the data volume by approximately 25%. Our sole test set was created using data from Rounds 1 and 2 and contained approximately 9000 segments.

### 3.2 Training

We began by training a base transformer model. We used the same early stopping strategy as in the En→De training and the same raw SentencePiece tokenization method. We then moved on to experiments using big transformer models. Here we also followed recommended settings for Marian as used in the En→De language pair. Our vocabulary size for both architectures in contrast was fixed for all cases at 36k. Our best result came from an ensemble of 4 models. The trainings were run on the same environments as the other language pairs, for less than 20 epochs for the base and 70 for the big transformer.

### 3.3 Unconstrained models

For the unconstrained models, the data initially included only some additional Euramis data related to the medical domain. To this data we tried to add segments from multiple sources — general OPUS and ParaCrawl data filtered on basic medical terms, the OPUS EMEA Corpus, and additional in house public health data. We again experimented with base and big transformer architectures, and again the best result came from a 4 model big transformer ensemble.

### 3.4 Results

| | | Test sets | |
|---|---|---|---|
| System | Data | R1+R2 (9k) | R2 off. (4k) |
| Base Tr. (c) | 900k | 51.8 | 20.3 |
| Big Tr. (c) | 900k | 54.3 | 20.0 |
| Big Tr. Ensemble (c) | 900k | 56.3 | **22.7** |
| Base Tr. Euramis (u) | 1.75M | 52.2 | 20.9 |
| Base Tr. multi (u) | 2.5M | 53.9 | 22.0 |
| Big Tr. multi ensemble (u) | 2.5M | 56.6 | **23.3** |

**Table 2.** Results for En→Sv models. *multi* is the multiple source data described above. R1+R2 is the development test set we created ourselves from data from Round 1 and Round 2. (c): constrained, (u): unconstrained model.

In Table 2 we present the BLEU scores for the main models we experimented with in the development of En→Sv systems. All models used filtered data. Despite the low number of segments in the training data (< 1M), using a big

transformer architecture here led to noticeably improved BLEU scores for both the constrained and unconstrained models. With our own test set, each addition of more data as well as passing from base to big to an ensemble architecture systematically increases the score. Results from the official test set show almost exactly the same behaviour apart from the base to big transformer where the trend is reversed. However, the big transformer ensemble shows an improvement of 2 points compared to the base architecture. The BLEU scores resulting from our own test set and the official test set are however vastly different. One reason may be that our own test set was filtered to some extent, though this alone would not explain such disparities.

## 4 English→Greek

### 4.1 Data

For the constrained task we used the data sets distributed for Round 1 and Round 2. For the unconstrained task training data from Euramis was added, similarly to the other language pairs. For both the constrained and unconstrained tasks we extracted validation and test data sets from Round 1 and Round 2 data. The number of segments in the data sets is summarized in Table 3.

| | Training | Validation | Testing |
|---|---|---|---|
| Constrained | 1,315,111 | 4,000 | 10,000 |
| Unconstrained | 2,259,617 | 4,000 | 10,000 |

**Table 3.** Number of segments used to train the En→El models.

There was no pre- or postprocessing of the data, although it was checked for sanity and consistency.

### 4.2 Models

For each of the two tasks (constrained and unconstrained) we trained (A) a base transformer (with default settings in the Marian toolkit) and (B) a big transformer (again with standard settings, see Section 2.2). We used a vocabulary size of 36k uniformly. Each model was trained on 4 GPUs.

### 4.3 Results

For En→El, we submitted the raw model output without any postprocessing. The BLEU scores are summarized in Table 4.

| | Test sets | |
|---|---|---|
| System | R1+R2 dev (10k) | R2 off. (4k) |
| Constrained Base Tr. | 47.9 | 41.7 |
| Constrained Big Tr. | 47.9 | 34.9 |
| Unconstrained Base Tr. | 48.3 | 44.3 |
| Unconstrained Big Tr. | 48.3 | 43.1 |

**Table 4.** Results for En→El models.

## 5 English→Spanish

The En→Es workflow was similar to the En→El one, except that the validation set was solely from Round 2 data. The data statistics are summarized in Table 5.

| | Training | Validation | Testing |
|---|---|---|---|
| Constrained | 3,428,760 | 4,000 | 10,000 |
| Unconstrained | 4,518,984 | 4,000 | 10,000 |

**Table 5.** Number of segments used to train the En→Es models.

### 5.1 Trainings and results

We experimented with the same setups as in En→El. Interestingly, our submitted systems scored much more competitively in this language pair than in En→El, our unconstrained system being the best by a significant margin. The results are summarized in Table 6.

| | Test sets | |
|---|---|---|
| System | R1+R2 dev (10k) | R2 off. (4k) |
| Constrained Base Tr. | 46.2 | 56.1 |
| Constrained Big Tr. | 46.6 | 56.1 |
| Unconstrained Base Tr. | 45.8 | 56.0 |
| Unconstrained Big Tr. | 46.4 | 56.5 |

**Table 6.** Results for En→Es models.

## 6 English→ Italian

### 6.1 Data

For the constrained task, we used the training data provided for Round 1 and Round 2. It was cleaned and filtered the same way as the data for En→De. The development set was used for validation and the test data was extracted at random from the cleaned and filtered training data (*dev_test_set*). A subset of the training data was extracted using a few keyword patterns, and this subset was used for fine-tuning.

For the unconstrained task, we experimented with adding data from the from Euramis and other sources, similarly to En→De. First, we extracted Euramis documents that, based on the metadata, were in-domain (*mtdata1*). The next addition were segments from the same database but extracted using keywords (*mtdata2*). The last addition was a combination of data from EMEA, TAUS Corona Crisis Corpora, and proprietary public-health data (*var_med*).

The number of segments in the training data is presented in Table 7.

| | Composition | Training | Validation | Testing |
|---|---|---|---|---|
| Constrained | R1+R2 | 1.6M | 4,000 | 10,000 |
| Constrained subset for FT | R1+R2 | 100k | 4,000 | 10,000 |
| Unconstrained | R1+R2+mtdata1 | 2.5M | 4,000 | 10,000 |
| Unconstrained | R1+R2+mtdata{1,2}+var_med | 3.7M | 4,000 | 10,000 |

**Table 7.** Number of segments used to train the En→It models. FT: fine-tuning.

### 6.2 Models and results

We built our baseline models for the constrained task as a standard base transformer model. Then we switched to a standard big transformer model. Training parameters were similar to En→De. Since the big transformer model gave better scores, we built 3 more for a 4 member ensemble. Finally, we fine-tuned all 4 big transformers.

For the unconstrained task, we experimented with a big transformer using *R1+R2+mtdata1*, a base transformer, and a 4 big transformer ensemble, using *R1+R2+mtdata1+mtdata2+var_med*. We did not normalize the translations, except in one case (cf. Table 8). The normalized translation got a slightly lower score.

We present the En→It results in Table 8. We did not expect that for such a small amount of training data (1.6–3.7M) the big transformer would be beneficial for the results but in this case it yielded better scores. According to the automatic evaluation, our unconstrained submission was the strongest system in this language pair.

| System | Data | Test sets R1+R2 dev (10k) | R2 off. (4k) |
|---|---|---|---|
| Base Tr. (c) | R1+R2 | 43.3 | 45.1 |
| Big Tr. (c) | R1+R2 | 46.3 | 44.0 |
| Big tr. ens.* (c) | R1+R2 | 47.7 | 47.0 |
| Big tr. ens. FT* (c) | R1+R2 | 47.8 | 46.7 |
| Big tr. (u) | R1+R2+mtdata1 | 46.3 | 46.6 |
| Base tr. (u) | R1+R2+mtdata1+mtdata2+var_med | 45.9 | 46.8 |
| Big tr. (u) | R1+R2+mtdata1+mtdata2+var_med | 48.5 | 48.3 |
| Big tr. ens.* (u) | R1+R2+mtdata1+mtdata2+var_med | 49.4 | 50.1 (49.9) |

**Table 8.** Results for En→It models. All scores are for non-normalized translations, except the score in parentheses; (c): constrained, (u): unconstrained model. Submissions are marked with an asterisk.

## 7 English→French

For the constrained En→Fr system we used the all the provided data filtered ...? Out unconstrained submissions were ...

Data is summarized in Table 9.

| | Training | Validation | Testing |
|---|---|---|---|
| Constrained | ... | ... | ... |
| Unconstrained | ... | ... | ... |

**Table 9.** Number of segments used to train the En→Fr models

### 7.1 Trainings and results

We experimented with the same setups as in ???. Our constrained systems were the winning submissions for this language pair outperforming the unconstrained submissions as well, while in the unconstrained submissions, it is worth noting that, similarly to En→De, the normalization of punctuation in postprocessing proved to make a significant (here 7 BLEU points!) difference. This might suggest that it would be beneficial to ensure some standardization in the reference sets with respect to typography to get a more reliable indication of the translation quality. The results are summarized in Table 10.

## 8 Conclusions

We described the submissions of the eTranslation team to the second round of the COVID19-MLIA translation shared task on 6 language pairs. Compared to

| | Test sets | |
|---|---|---|
| System | Dev (?k) | R2 off. (4k) |
| Constrained 1? | ... | 57.9 |
| Constrained 2? | ... | 58.3 |
| Unconstrained normalized | ... | 49.9 |
| Unconstrained raw | ... | 56.9 |

**Table 10.** Results for En→Fr models.

Round 1, our systems were more competitive ending up in first place in several categories. We tried to focus on data selection and filtering and experimented with some complex architectures, and automatic evaluation results justified this approach. We hypothesize that when similar development workflows resulted in worse positions in the rankings for different language pairs, the diversity and noise in the data sets might have played a role: some of our models performed accidentally better or worse on the specific datasets but the general performance of these models are most likely similar. This hypothesis has some support from the strong performance of the most powerful unconstrained systems, which is probably due to their robustness. However, further testing would be necessary to confirm these assumptions.

## References


[1] Biljon, E.V., Pretorius, A., Kreutzer, J.: On optimal transformer depth for low-resource language translation. CoRR **abs/2004.04418** (2020), URL `https://arxiv.org/abs/2004.04418`

[2] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016), URL `https://arxiv.org/abs/1612.03651`

[3] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations, pp. 116–121, Association for Computational Linguistics (2018), URL `http://aclweb.org/anthology/P18-4020`

[4] Kudo, T.: 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), pp. 66–75, Association for Computational Linguistics (2018), URL `http://aclweb.org/anthology/P18-1007`

[5] Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., Gilbro, S.: An overview of the European Union's highly multilingual parallel corpora. Language Resources and Evaluation **48**(4),

679–707 (Dec 2014), ISSN 1574-0218, https://doi.org/10.1007/s10579-014-9277-0, URL https://doi.org/10.1007/s10579-014-9277-0