# Findings of a Machine Translation Shared Task Focused on Covid-19 Related Documents

Francisco Casacuberta[1,2], Alexandru Ceausu[3], Khalid Choukri[4], Miltos Deligiannis[5], Miguel Domingo[1,2], Mercedes García-Martínez[6], Manuel Herranz[6], Guillaume Jacquet[7], Vassilis Papavassiliou[5], Stelios Piperidis[5], Prokopis Prokopidis[5], Dimitris Roussis[5] and Marwa Hadj Salah[4]

[1] *PRHLT Research Center, Universitat Politècnica de València*
[2] *ValgrAI - Valencian Graduate School and Research Network for Artificial Intelligence*
[3] *Court of Justice of the EU*
[4] *Evaluations and Language resources Distribution Agency*
[5] *Athena Research Center*
[6] *Pangeanic S. L. - PangeaMT*
[7] *European Commission*

Abstract
This work presents the results of a machine translation shared task focused on Covid-19 related documents. Nine teams took part in this event, which was divided in two rounds and involved seven different language pairs. Two different scenarios were considered: one in which only the provided data was allowed, and a second one in which the use of external resources was allowed. Overall, the best approaches were based on multilingual models and transfer learning, with an emphasis on the importance of applying a cleaning process to the training data.

**Keywords**
Machine Translation, Evaluation, Shared Task, Covid-19

## 1. Introduction

In emergency situations, the public as well as many other stakeholders need to aggregate and summarize different sources of information into a single coherent synopsis or narrative, complementing different pieces of information, resolving possible inconsistencies and preventing misinformation. This should happen across multiple languages, sources and levels of linguistic knowledge that varies depending on social, cultural or educational factors.

As a response to the Covid-19 crisis, the Covid-19 MLIA initiative[1] organized a community evaluation effort aimed at accelerating the creation of resources and tools for improving the deployment of automatic systems focused on Covid-19 related documents. This initiative consisted of three natural language processing tasks: (1) information extraction, (2) multilingual semantic search and (3) machine translation.

In this paper, we focus on the third task about machine translation (MT), which was focused on texts from the Covid-19 crisis that shocked the world and for which there were not many processed text or corpora. The task was divided in two rounds. At the end of each round, participants wrote or updated their report describing their system and highlighting which methods and data had been used.

## 2. Task description

The goal of this shared task was to benchmark MT systems focused on Covid-19 related documents for several language pairs. Fig. 1 shows some examples of sentences from Covid-19 related documents. A total of 7 different language pairs were addressed throughout the initiative: English–German, English–French, English–Spanish, English–Italian, English–Modern Greek, English–Swedish and English–Arabic (second round only).

Given a set of training data provided by the organizers for each language pair, participants had to train up to five different MT systems per language pair. These systems were classified in two scenarios:

- **Constrained**: systems could be trained exclu-

*SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.*
fcn@prhlt.upv.es (F. Casacuberta); Alexandru.Ceausu@curia.europa.eu (A. Ceausu); choukri@elda.org (K. Choukri); mdel@athenarc.gr (M. Deligiannis); midobal@prhlt.upv.es (M. Domingo); m.garcia@pangeanic.com (M. García-Martínez); m.herranz@pangeanic.com (M. Herranz); guillaume.jacquet@ec.europa.eu (G. Jacquet); vpapa@athenarc.gr (V. Papavassiliou); spip@athenarc.gr (S. Piperidis); prokopis@athenarc.gr (P. Prokopidis); dimitris.roussis@athenarc.gr (D. Roussis); marwa@elda.org (M. H. Salah)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] http://eval.covid19-mlia.eu/.

*30% of children and adults infected with measles can develop complications.*
*The first dose is given between 10 and 18 months of age in European countries.*

**Figure 1:** Examples of English sentences from Covid-19 related documents.

sively with data provided by the organizers (including data from a different language pair, monolingual data, etc). The use of basic linguistic tools such as taggers, parsers or morphological analyzers or multilingual systems was allowed for this scenario.

- **Unconstrained**: systems could be trained using data not provided by the organizers, from any external resource not allowed in the constrained scenario.

Systems were evaluated and compared according to the scenario to which they belonged. It was mandatory that one of the submitted systems per language pair belonged to the constrained scenario. Participants were able to take part in any or all of the language pairs. They used their systems to translate a test set of unseen sentences in the source language. Evaluation consisted on assessing the translation quality of the submissions. Different metrics were used on each round.

## 2.1. Data collection

In the context of the first round of this initiative, we decided to collect an initial collection of parallel corpora in health and medicine domains from well-known web sources and enrich them with identified Covid-19 parallel data. The purpose of following this approach was to implement a very quick response of the MT community in an emergency situation, like the current pandemic.

To this end, we first collected an updated version of the European Medicines Agency (EMEA) corpus[2], and applied new (more robust and efficient) methods for text extraction from *pdf* files, sentence splitting, sentence alignment and parallel corpus filtering. Moreover, medical-related multilingual collections which were offered by the Publications Office of EU[3] were processed in a similar manner, increasing the volume of the "general" subset of the training data.

The first step of acquiring Covid-19-related data was the identification of several bilingual websites with such content. With the aim of constructing datasets that could be publicly available, we targeted websites of national authorities and public health agencies[4], EU agencies and specific broadcast websites (e.g., voxeurop[5], GlobalVoices[6] or Voltairnet[7]).

For acquiring domain-specific bilingual corpora, we used a recent version of ILSP-FC [1], a modular toolkit that integrates modules for text normalization, language identification, document clean-up, text classification, bilingual document alignment (i.e., identification of pairs of documents that are translations of each other) and sentence alignment. As mentioned above, taking into account the emergency situation, a "rapid" approach based on keywords was adopted for text classification (i.e., keeping only documents that are strongly related to the current worldwide health crisis). Specifically for sentence alignment, the LASER[8] toolkit was used instead of the integrated aligner. Then, a battery of criteria was applied on aligned sentences to automatically filter out sentence pairs with potential alignment or translation issues (e.g., with score less than a predefined threshold) or of limited use for training MT systems (e.g., duplicate pairs, identical segments in a pair, etc.) and, thus, generate precision-high language resources.

For the second round, we repeated the previous process—re-crawling several websites of national authorities and public health agencies—in order to enrich the data that had already been collected. Additionally, we exploited the outcomes of an available infrastructure, namely Medical Information System (MediSys[9]), with the purpose of constructing parallel corpora beneficial for MT [2]. Similarly to the first round's approach, it could be seen as an application of implementing a quick response of the MT community to the pandemic crisis.

MediSys is one of the publicly accessible systems of the Europe Media Monitor (EMM) which processes media to identify potential public health threats in a fully automated fashion [3]. Focusing on the current pandemic, a dataset of metadata which concerns Covid-19 related news was made publicly available in RSS/XML format, which corresponded to millions of news articles [4]. The dataset was divided into subsets according to the articles' month of publication. First, the metadata were parsed and the URL and language of each article were extracted. Then, each web page of the targeted languages was fetched and its main content was stored in a text file. The generated text files were merged to create a single document for each language and each period. Thus, these documents constituted the Covid-19 related monolingual corpora and

[2] https://www.prhlt.upv.es/~mt/prokopidis-and-papavassiliou-emea.html.
[3] https://op.europa.eu/en/home.
[4] Such a list is available at https://www.ecdc.europa.eu/en/COVID-19/national-sources.
[5] https://voxeurop.eu/.
[6] https://globalvoices.org/.
[7] https://www.voltairenet.org/.
[8] https://github.com/facebookresearch/LASER.
[9] https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/LANGUAGE-TECHNOLOGY/EMM_collection/2020_MediSys_Covid19_dataset/.

were considered comparable (in pairs), due to their narrow topic and the fact that they were published in the same time period. To this end, the LASER toolkit was applied on each document pair to mine sentence alignments for each EN-X language pair. Finally, several filtering methods were adopted (i.e., thresholding the alignment score by 1.04, removing near de-duplicates, etc) to compile the final dataset.

## 2.2. Corpora

For the first round, we selected the data described in the previous section (Section 2.1) and split them into train, validation and test. Then, to ensure that the tests were a good representation of the task and were appropriate for being used for evaluation, we sorted all segments from the initial test according to the alignment probability between source and target. After that, we filtered them according to their number of words: removing those segments whose source had either less than 0.7 or more than 1.3 times the average number of words per sentence from the training set. Finally, we selected the first two thousand segments to construct the final version of the test set for round 1.

This process was improved for selecting round 2's corpora. Given the data used for this round, we computed some statistics and removed the outliers (segments that contained more than 100 words in either its source or target). Then, we split the data into train, validation and test sets. Since the data came from different sources, we wanted to ensure that both the validation and tests sets were representative enough of the training sets. For this reason, for each language pair, we computed the representation of each source in the total data (i.e., the number of segments from this source divided by the total number of segments). Then, out of the total segments we wanted to select for validation and test (4000 for each), we select that same percentage from each source.

Additionally, to ensure that validation and test did not contain low-quality segments (given that the data had been crawled from the web), we sorted the segments according to its alignment quality. Finally, we shuffled the selected segments and split them equally into validation and test.

Therefore, the procedure we followed for each language pair was:

1. We computed the ratio of data from each different source over the total data.
2. We computed the average number of words per segment over this set.
3. We constituted a subset [0.7 * average words per segment, 1.3 * average words per segment].
4. We sorted this subset (from best to worst) according to its alignment score.
5. We selected the best 8000 * the percentage obtained at step 1 segments.
6. We shuffled those segments and select half of them for validation and the other half for test.

Table 1 describes the corpora statistics.

## 2.3. Quality assessment

Taking into account that the corpora were obtained from crawling (see Section 2.1), it is important to assess the quality of the reference sets. To do so, we selected a subset of the Spanish round 1 corpora and post-edited it with the help of a team of professional translators. This subset consisted of the worst 500 segments according to the alignment probability between source and reference. Overall, translators thought that *the translations in general are good, but some are very free, adding things that are not in the source, or they are too literal*.

To assess the quality of the reference sets, we compared the reference and its post-edited version using human TER (hTER) [5]. This metric computes the number of errors between a translation hypothesis and its post-edited version (in this case, between the automatic reference and its post-edited version). Thus, the smallest the value the highest the quality. We obtained a fairly low hTER value (18.8), which indicates that the translation quality of the reference is generally good and, thus, is coherent with the translators' opinion.

## 2.4. Evaluation

In order to evaluate the participant's systems, we selected the bilingual evaluation understudy (BLEU) [6]—which computes the geometric average of the modified n-gram precision, multiplied by a brevity factor—as our main metric, using `sacreBLEU` [7] to compute it ensuring consistent scores.

Additionally, we selected different alternative well-known MT metrics for each round:

- Round 1:
  - **Character n-gram F-score (ChrF)** [8]: character n-gram precision and recall arithmetically averaged over all character n-grams.
- Round 2:
  - **Translation Edit Rate (TER)** [5]: this metric computes the number of word edit operations (insertion, substitution, deletion and swapping), normalized by the number of words in the final translation.
  - **BEtter Evaluation as Ranking (BEER)** [9]: a sentence level metric that incorporates a large number of features combined in a linear model.

**Table 1**
Corpora statistics, divided by rounds. $|S|$ stands for number of sentences, $|T|$ for number of tokens and $|V|$ for size of the vocabulary. M denotes millions and K thousands.

**Round 1**

| | | German | | French | | Spanish | | Italian | | Modern Greek | | Swedish | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | En | De | En | Fr | En | Es | En | It | En | El | En | Sv |
| Train | $\|S\|$ | 926.6K | | 1.0M | | 1.0M | | 900.9K | | 834.2K | | 806.9K | |
| | $\|T\|$ | 17.3M | 16.1M | 19.4M | 22.6M | 19.5M | 22.3M | 16.7M | 18.2M | 15.0M | 16.4M | 14.5M | 13.2M |
| | $\|V\|$ | 372.2K | 581.6K | 401.0K | 438.9K | 404.4K | 458.0K | 347.7K | 416.0K | 305.7K | 407.5K | 298.2K | 452.0K |
| Validation | $\|S\|$ | 528 | | 728 | | 2.5K | | 3.7K | | 3.9K | | 723 | |
| | $\|T\|$ | 8.2K | 7.6K | 17.0K | 18.8K | 48.9K | 56.2K | 78.2K | 84.0K | 73.0K | 72.7K | 11.4K | 10.0K |
| | $\|V\|$ | 2.4K | 2.6K | 4.1K | 4.5K | 9.7K | 10.6K | 12.4K | 14.9K | 10.3K | 14.5K | 2.6K | 2.8K |
| Test | $\|S\|$ | 2000 | | 2000 | | 2000 | | 2000 | | 2000 | | 2000 | |
| | $\|T\|$ | 34.9K | 33.2K | 33.2K | 35.8K | 32.6K | 34.3K | 33.7K | 34.2K | 42.6K | 44.3K | 35.3K | 30.6K |
| | $\|V\|$ | 7.8K | 9.6K | 6.7K | 7.7K | 6.7K | 7.9K | 8.6K | 10.4K | 9.5K | 12.5K | 7.1K | 8.2K |

**Round 2**

| | | German | | French | | Spanish | | Italian | | Modern Greek | | Swedish | | Arabic | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | En | De | En | Fr | En | Es | En | It | En | El | En | Sv | En | Ar |
| Train | $\|S\|$ | 1.5M | | 2.4M | | 2.9M | | 1.0M | | 674.0K | | 375.0K | | 424.4K | |
| | $\|T\|$ | 23.5M | 22.1M | 45.6M | 53.0M | 52.4M | 60.3M | 16.4M | 17.2M | 11.4M | 12.2M | 5.5M | 5.1M | 7.7M | 7.5M |
| | $\|V\|$ | 523.9K | 847.5K | 782.2K | 781.4K | 850.0K | 950.2K | 421.2K | 501.3K | 289.7K | 378.7K | 180.7K | 234.7K | 222.2K | 360.2K |
| Validation | $\|S\|$ | 4.0K | | 4.0K | | 4.0K | | 4.0K | | 4.0K | | 4.0K | | 4.0K | |
| | $\|T\|$ | 62.2K | 61.2K | 72.0K | 83.9K | 72.2K | 81.4K | 64.6K | 69.0K | 67.8K | 72.5K | 56.6K | 54.4K | 75.9K | 74.7K |
| | $\|V\|$ | 13.9K | 17.1K | 13.2K | 14.8K | 13.8K | 15.8K | 14.6K | 16.7K | 14.0K | 18.0K | 12.3K | 14.1K | 16.1K | 23.7K |
| Test | $\|S\|$ | 4.0K | | 4.0K | | 4.0K | | 4.0K | | 4.0K | | 4.0K | | 4.0K | |
| | $\|T\|$ | 62.2K | 61.0K | 72.3K | 84.1K | 72.2K | 81.4K | 64.3K | 68.7K | 67.8K | 72.4K | 56.5K | 54.3K | 76.1K | 74.5K |
| | $\|V\|$ | 13.8K | 17.0K | 13.1K | 14.8K | 13.7K | 15.7K | 14.4K | 16.7K | 14.1K | 18.2K | 12.3K | 14.1K | 16.2K | 23.5K |

We applied approximate randomization testing (ART) [10]—with $10,000$ repetitions and using a $p$-value of $0.05$—to determine whether two systems presented statistically significance. The scripts used for conducting the automatic evaluation are publicly available together with some utilities which were useful for the shared task[10].

Following the WMT criteria [11], we grouped systems together into clusters according to the statistical significance of their performance (as determined by ART). With that purpose, we sorted the submissions according to each metric and computed the significance of the performance between one system and the following. If it was not significant, we added the second system into the cluster of the first system[11]. Otherwise, we added it into a new cluster. This way, systems from one cluster significantly outperformed all others in lower ranking clusters.

## 2.5. Baselines

At each round, we trained two different constrained systems to use as baselines in order to have an estimation of the expected translation quality of each scenario. The first system was based on recurrent neural network (RNN) [12, 13] while the other one was based on the Transformer architecture [14]. All systems were built using `OpenNMT-py` [15].

[10]Hidden GitHub repository.

[11]Considering that, at the start of this process, there is an initial cluster containing the first system.

### RNN

These systems were trained using the standard parameters for RNN MT systems: long short-term memory units [16], with all model dimensions set to $512$; Adam [17], with a fixed learning rate of $0.0002$ and a batch size of $60$; label smoothing of $0.1$ [18]; beam search with a beam size of $6$; and joint byte pair encoding (BPE) [19] applied to all corpora, using $32,000$ merge operations. In light of the results, this architecture was only used for the first round.

### Transformer

These systems were trained using the standard parameters: $6$ layers; Transformer [14], with all dimensions set to $512$ except for the hidden transformer feed-forward (which was set to $2048$); $8$ heads of Transformer self-attention; $2$ batches of words in a sequence to run the generator on in parallel; a dropout of $0.1$; Adam [17], using an Adam beta2 of $0.998$, a learning rate of $2$ and Noam learning rate decay with $8000$ warm up steps; label smoothing of $0.1$ [18]; beam search with a beam size of $6$; and joint BPE applied to all corpora, using $32,000$ merge operations.

## 2.6. Participants' approaches

In this subsection, we present the different approaches submitted by each team.

### Accenture

This team only participated in round 1, with an approach based on multilingual BART [20].

### CdT-ASL

CdT-ASL team developed NICE which integrates neural machine translation (NMT) custom engines for confidential adapted translations. They submitted constrained and unconstrained systems, they added generic and public health domains to internal data for unconstrained systems. They applied cleaning processes to prepare the data for training with big transformer using *OpenNMT-tf*. They only took part in round 2.

### CUNI-MT

This team took part in both rounds. For the first round, they submitted approaches based on standard NMT with online back-translation [21]; a transfer learning approach based on Kocmi and Bojar [22]; and multilingual models in which, during inference, the corresponding embedding of the target language was selected. For the second round, they trained a multilingual model jointly on all languages.

### CUNI-MTIR

CUNI-MTIR only took part in round 1, training constrain models using the Transformer architecture from MarianNMT [23], and using the UFAL Medical Corpus[12] for training unconstrained data and then fine-tuning the models with the constrained data. All the data was tokenized using Khresmoi[13]'s tokenizer and, then, encoded using BPE with 32K merges.

### E-Translation

For round 1, this team used transfer learning and a 12K size vocabulary created using SentencePiece over Transformer models trained with MarianNMT [23]. Additionally, they submitted to the unconstrained category their WMT system and a new version of that system, fine-tuned with the constrained data.

For round 2, they focused on performing a general clean-up including a language identifier and checking the match of the number of tokens in source and target to filter noisy segments. For Greek and Spanish they did not do pre- or post-processing, only sanity checking. They experimented with standard Transformer and big Transformer in MarianNMT [23]. For the unconstrained scenario, they made use of the TAUS Corona Crisis Corpora, the OPUS EMEA Corpus and a health related subset of the Euramis dataset.

### Lingua Custodia

Lingua Custodia's submissions for round 1 consisted of a multilingual model able to translate from English to French, German, Spanish, Italian and Swedish; and individual translation models for English–German and English–French. They applied unigram SentencePiece for subword segmentation using a source and target shared vocabulary of 50K for individual models and 70K for multilingual models. Additionally, authors split the numbers character by character. For multilingual models, a language token is added to the source in order to indicate the target language. The English–German multilingual model achieved much higher score than the English–German single model. This improvement is not shown in the English–French model.

For round 2, they participated in the constrained scenario. The pre-processing used was based on Moses' tokenizer and cleaning techniques such as removing much longer sentences comparing source and target lengths, replacement of consecutive spaces by one space. They used inline casing consisting of adding a tag with the casing information. They finally append the language token to each source sentence in the pre-processing in order to indicate the target language for multilingual models. Standard transformer architecture in Sockeye toolkit was used for training in multiple GPUs instead of Seq2SeqPy used previously because the data loading is more efficient and has better support for multiple GPUs.

### LIMSI

LIMSI only took part in round 1, submitting a Transformer model using BPE with 32K vocabulary units was applied to the constrained system. They submitted four unconstrained systems: 1) one system build using an external in-domain biomedical corpora; 2) a system first trained on WMT14[14] general data and fine-tuned on the shared task's corpus; 3) same as 2) but adding BERT [24]; and 4) a system only trained with constrained data but computing the BPE codes using all the external in-domain corpus.

### PROMT

For round 1, PROMT's approaches consisted in a multilingual model trained using MarianNMT's [23] Transformer architecture. For the constrained scenario, all data was concatenated using de-duplication to one single multilingual corpus to build a 8k SentencePiece [25] model for subword segmentation. In addition, a language-specific tag was added to the source side of the parallel sentence

[12] http://ufal.mff.cuni.cz/ufal_medical_corpus.

[13] http://www.khresmoi.eu/assets/Deliverables/WP4/KhresmoiD412.pdf.

[14] http://www.statmt.org/wmt14/translation-task.html.

**Table 2**

Results of the first round, divided by categories. Systems are ranked according to BLEU. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

**English–German**

| | Rank | Team | Description | BLEU [↑] | chrF [↑] |
|---|---|---|---|---|---|
| Constrained | 1 | CUNI-MT | transfer2 | 31.6 | 0.600 |
| | | CUNI-MT | base | 31.4 | 0.596 |
| | | CUNI-MT | transfer1 | 31.3 | 0.595 |
| | | PROMT | multilingual | 31.1 | 0.599 |
| | 2 | E-Translation | basetr | 30.4 | 0.593 |
| | 3 | CUNI-MT | transfer3 | 29.8 | 0.584 |
| | | Lingua Custodia | multilingual | 29.5 | 0.584 |
| | 4 | Baseline | Transformer | 28.1 | 0.573 |
| | | Lingua Custodia | transformer | 26.7 | 0.556 |
| | 5 | TARJAMA-AI | base3 | 25.6 | 0.564 |
| | 6 | TARJAMA-AI | base2 | 25.0 | 0.559 |
| | 7 | CUNI-MTIR | r1 | 19.7 | 0.494 |
| | 8 | Baseline | RNN | 17.9 | 0.479 |
| | | TARJAMA-AI | base | 17.7 | 0.488 |
| Unconstrained | 1 | E-Translation | wmtfinetune | 44.4 | 0.686 |
| | 2 | E-Translation | wmt | 44.1 | 0.683 |
| | 3 | PROMT | transformer | 41.2 | 0.666 |
| | 4 | CUNI-MTIR | r1 | 20.0 | 0.499 |

**English–French**

| | Rank | Team | Description | BLEU [↑] | chrF [↑] |
|---|---|---|---|---|---|
| Constrained | 1 | PROMT | multilingual | 49.6 | 0.711 |
| | 2 | E-Translation | small | 49.1 | 0.707 |
| | | Lingua Custodia | multilingual | 49.0 | 0.705 |
| | | Lingua Custodia | transformer | 48.9 | 0.703 |
| | | CUNI-MT | base | 48.4 | 0.703 |
| | | CUNI-MT | multiling | 48.0 | 0.700 |
| | | E-Translation | big | 47.4 | 0.695 |
| | | Baseline | Transformer | 47.3 | 0.693 |
| | | CUNI-MT | transfer2 | 47.1 | 0.693 |
| | 3 | LIMSI | trans | 43.5 | 0.660 |
| | 4 | CUNI-MTIR | r1 | 34.9 | 0.605 |
| | - | Baseline | RNN | 34.3 | 0.596 |
| | 5 | TARJAMA-AI | base | 26.8 | 0.567 |
| | 6 | Accenture | mbart | 15.8 | 0.464 |
| Unconstrained | 1 | PROMT | transformer | 59.5 | 0.767 |
| | 2 | E-Translation | gen | 52.9 | 0.742 |
| | 3 | LIMSI | indom | 51.2 | 0.721 |
| | 4 | E-Translation | phwt | 50.1 | 0.724 |
| | | LIMSI | trans | 49.3 | 0.710 |
| | | LIMSI | bert | 49.3 | 0.703 |
| | | LIMSI | mlia | 48.5 | 0.705 |
| | 5 | E-Translation | eufl | 47.9 | 0.712 |
| | 6 | CUNI-MTIR | r1 | 33.0 | 0.590 |

**English–Spanish**

| | Rank | Team | Description | BLEU [↑] | chrF [↑] |
|---|---|---|---|---|---|
| Constrained | 1 | PROMT | multilingual | 48.3 | 0.702 |
| | 2 | CUNI-MT | transfer1 | 47.9 | 0.699 |
| | | CUNI-MT | transfer2 | 47.6 | 0.698 |
| | | Lingua Custodia | multilingual | 47.5 | 0.695 |
| | | Baseline | Transformer | 47.4 | 0.694 |
| | | CUNI-MT | multiling | 47.3 | 0.692 |
| | | CUNI-MT | base | 47.3 | 0.691 |
| | - | Baseline | RNN | 35.6 | 0.609 |
| | 3 | CUNI-MTIR | r1 | 32.9 | 0.591 |
| | 4 | TARJAMA-AI | base | 30.9 | 0.593 |
| | 5 | Accenture | mbart | 17.4 | 0.474 |
| Uncon. | 1 | PROMT | transformer | 58.2 | 0.762 |
| | 2 | CUNI-MTIR | r1 | 32.1 | 0.582 |

**English–Italian**

| | Rank | Team | Description | BLEU [↑] | chrF [↑] |
|---|---|---|---|---|---|
| Constrained | 1 | PROMT | multilingual | 29.6 | 0.585 |
| | 2 | Lingua Custodia | multilingual | 28.4 | 0.572 |
| | | CUNI-MT | transfer2 | 28.3 | 0.574 |
| | | CUNI-MT | multiling | 28.3 | 0.574 |
| | - | Baseline | Transformer | 26.9 | 0.560 |
| | 3 | TARJAMA-AI | base | 19.2 | 0.494 |
| | - | Baseline | RNN | 17.0 | 0.473 |
| Uncon. | 1 | PROMT | transformer | 38.0 | 0.642 |

**English–Modern Greek**

| | Rank | Team | Description | BLEU [↑] | chrF [↑] |
|---|---|---|---|---|---|
| Constrained | 1 | PROMT | multilingual | 27.2 | 0.523 |
| | 2 | CUNI-MT | transfer1 | 24.7 | 0.496 |
| | 3 | CUNI-MT | base | 24.1 | 0.484 |
| | | Baseline | Transformer | 22.6 | 0.471 |
| | - | Baseline | RNN | 12.8 | 0.365 |
| Uncon. | 1 | PROMT | transformer | 42.4 | 0.652 |

**English–Swedish**

| | Rank | Team | Description | BLEU [↑] | chrF [↑] |
|---|---|---|---|---|---|
| Constrained | 1 | PROMT | multilingual | 30.7 | 0.595 |
| | | Lingua Custodia | multilingual | 30.4 | 0.589 |
| | | CUNI-MT | transfer2 | 30.1 | 0.590 |
| | 2 | CUNI-MT | transfer | 28.5 | 0.578 |
| | - | Baseline | Transformer | 27.8 | 0.566 |
| | 3 | CUNI-MT | base | 26.6 | 0.561 |
| | 4 | CUNI-MTIR | r1 | 25.1 | 0.541 |
| | - | Baseline | RNN | 19.2 | 0.481 |
| | 5 | TARJAMA-AI | base | 11.2 | 0.443 |
| Uncon. | 1 | PROMT | transformer | 41.3 | 0.671 |
| | 2 | CUNI-MTIR | r1 | 24.0 | 0.514 |

pairs (e.g., $< it >$ token was added to the beginning of the English sentence of the English–Italian sentence pair). They also removed all tokens that appeared less than ten times in the combined de-duplicated monolingual corpus from their vocabulary.

For the unconstrained scenario, all available data mainly from the OPUS [26] and statmt[15] with the addition of private data harvested from the Internet were added to the training data. A special BPE implementation [27] developed by the team was applied instead of SentencePiece, but the authors used SentencePiece in the constrained scenario as it seemed to work better in low-resource settings. The size of the BPE models and vocabularies varied from 8k to 16k and shared vocabulary was not used (separate BPE models were trained) for the English–Greek pair as the two languages have different alphabets.

For round 2, they trained a transformer multilingual model with a single encoder and a single decoder with Marian toolkit and performing fine-tuning for each language pair. For the unconstrained scenario, they used the same approach as in round 1.

#### TARJAMA-AI

This team submitted a single system consisting in a model trained with all the language pairs data adding a special token for the non-target languages. Additionally, they over-sampled the corpus of the desired target language (i.e., the English–Spanish corpus for training the constrained English–Spanish, etc). They only took part in round 1.

## 3. Results

In this section, we present the results from each round. Following the WMT criteria [11], we grouped systems together into clusters according to which systems significantly outperformed all others in lower ranking clusters, according to ART (see Section 2.4). For clarity purposes and space constrains, we use BLEU as the main metric for performing the ranking. Nonetheless, we tried using each metric from Section 2.4 as the main one for ranking, observing that all of them resulted in similar clusters.

### 3.1. Round 1

Table 2 presents the results of the first round. Overall, *multilingual* and *transfer learning* approaches yielded the best results for all language pairs in the constrained scenario. In fact, except for English–German (in which they shared the same ranking), *PROMT*'s multilingual approach—which was the only multilingual system trained for all language pairs—achieved the best results in all cases. This approach also used a smaller vocabulary and *SentencePiece* instead of BPE.

In general, the differences from one position to the next one were of a few points (according to both metrics), with a case (English–French) in which there are two points of difference (according to BLEU) between the first and last approaches of the same ranking. Our baselines worked well as delimiters: more sophisticated approaches generally ranked above our Transformer baselines, while the rest ranked either between them or below the RNN baselines. Moreover, the RNN baselines established the limit before a significant drop in translation quality between approaches of one position in the ranking and the next position (sometimes it is the exact limit, while other times there is a cluster above it of a similar quality).

Regarding the unconstrained scenario, it had less participation than the constrained one. With an exemption (*E-Translation*'s approaches based on their WMT system [28] yielded the best results for English–German), *PROMT*'s multilingual approach achieved the best results for all language pairs. In general, approaches were similar to the constrained ones but using additional external data. Additionally, due to the use of external data, the best unconstrained systems yielded around 10 BLEU points and 7 ChrF points of improvement compared to the best constrained systems for each language pair.

### 3.2. Round 2

Table 3 presents the results of the second round. With the exception of English–French, in which *monolingual approaches* achieved the best results, *multilingual approaches* yielded the best performances. In the case of English—German, *system ensembling* also ranked at first position.

In general, the differences from one position to the next one were of a few points (according to all metrics). Our baselines worked well as delimiters: more sophisticated approaches generally ranked above our baselines, while the following cluster after them obtained the highest quality drop between consecutive ranks.

## 4. Conclusions

This work presents a community evaluation effort to improve the generation of MT systems as a response to a global problem. The initiative consisted of generating specialized corpus for a new and important topic: Covid-19. This initiative was divided into two rounds.

This first round addressed 6 different language pairs and was divided into two scenarios: one in which participants were limited to using only the provided corpora (constrained) and another one in which the use of external tools and data was allowed (unconstrained). 8 different

[15] http://www.statmt.org/.

**Table 3**
Results of the second round, divided by categories. Systems are ranked according to BLEU. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally. The baseline *Transformer+* corresponds to the one trained using also the data from round 1.

**English–German**

| | Rank | Team | Description | BLEU [↑] | TER [↓] | BEER [↑] |
|---|---|---|---|---|---|---|
| Constrained | 1 | Lingua Custodia | 5lang-ft-avg | 40.3 | 48.4 | 66.8 |
| | | E-Translation | ensembleFT | 39.9 | 48.2 | 66.8 |
| | | Lingua Custodia | 5lang-ft | 39.8 | 48.9 | 66.5 |
| | | E-Translation | ensemble | 39.7 | 48.4 | 66.6 |
| | | Lingua Custodia | 1lang | 39.7 | 50.1 | 65.9 |
| | 2 | PROMT | multilingual-model-round2-tuned-de | 39.6 | 47.7 | 66.8 |
| | 3 | Lingua Custodia | 7lang | 38.6 | 50.0 | 65.8 |
| | 4 | PROMT | multilingual-model-round2 | 39.6 | 49.6 | 65.7 |
| | - | Baseline | Transformer | 34.9 | 51.7 | 63.9 |
| | | Baseline | Transformer+ | 34.8 | 51.8 | 63.7 |
| | 5 | CUNI-MT | transfer | 31.8 | 54.6 | 61.8 |
| | 6 | PROMT | multilingual-model-round1 | 28.7 | 57.7 | 60.8 |
| | 7 | CUNI-MT | transfer2 | 27.5 | 60.4 | 59.8 |
| | 8 | CUNI-MT | multiling | 27.0 | 60.9 | 59.2 |
| Unconstrained | 1 | E-Translation | wmtFT | 45.7 | 43.0 | 70.4 |
| | 2 | PROMT | Transformer | 40.4 | 46.9 | 67.9 |
| | 3 | E-Translation | singlebigTr | 40.0 | 48.4 | 66.9 |
| | 4 | E-Translation | eTstandardengine | 35.4 | 52.7 | 64.6 |

**English–French**

| | Rank | Team | Description | BLEU [↑] | TER [↓] | BEER [↑] |
|---|---|---|---|---|---|---|
| Constrained | 1 | E-Translation | 2 | 58.3 | 33.8 | 75.1 |
| | | E-Translation | 1 | 57.9 | 34.0 | 75.0 |
| | | Lingua Custodia | 1lang | 57.2 | 34.9 | 74.5 |
| | | PROMT | multilingual-model-round2-tuned-fr | 57.1 | 34.1 | 74.8 |
| | 2 | CdT-ASL | only-round2-data | 56.9 | 34.6 | 74.5 |
| | 3 | Lingua Custodia | 7lang | 55.8 | 35.7 | 73.9 |
| | | PROMT | multilingual-model-round2 | 55.4 | 35.2 | 74.0 |
| | - | Baseline | Transformer | 54.4 | 35.9 | 73.4 |
| | - | Baseline | Transformer+ | 53.7 | 36.7 | 73.0 |
| | 4 | PROMT | multilingual-model-round1 | 45.4 | 43.1 | 68.9 |
| | 5 | CUNI-MT | multiling | 44.1 | 45.6 | 67.6 |
| Unconstrained | 1 | PROMT | Transformer | 57.1 | 34.5 | 74.8 |
| | | E-Translation | generaldenorm | 56.9 | 34.8 | 74.5 |
| | 2 | E-Translation | general | 49.9 | 38.8 | 72.0 |
| | | CdT-ASL | only-cdt-data | 49.7 | 40.0 | 71.3 |
| | 3 | E-Translation | formal | 43.5 | 44.6 | 68.0 |

**English–Spanish**

| | Rank | Team | Description | BLEU [↑] | TER [↓] | BEER [↑] |
|---|---|---|---|---|---|---|
| Constrained | 1 | Lingua Custodia | 1lang-avg | 56.6 | 33.7 | 75.2 |
| | | E-Translation | 2 | 56.1 | 33.5 | 75.2 |
| | | E-Translation | 1 | 56.1 | 33.5 | 75.2 |
| | | Lingua Custodia | 5lang-ft-avg | 56.0 | 33.8 | 75.1 |
| | 2 | CdT-ASL | only-round2-data | 55.4 | 34.1 | 74.6 |
| | | Lingua Custodia | 7lang | 55.3 | 34.4 | 74.8 |
| | | PROMT | multilingual-model-round2-tuned-es | 54.9 | 33.9 | 74.9 |
| | 3 | PROMT | multilingual-model-round2 | 53.8 | 34.5 | 74.3 |
| | - | Baseline | Transformer | 53.3 | 35.2 | 74.0 |
| | - | Baseline | Transformer+ | 51.8 | 36.1 | 73.3 |
| | 4 | CUNI-MT | transfer | 48.4 | 39.3 | 71.2 |
| | 5 | PROMT | multilingual-model-round1 | 45.1 | 41.2 | 69.9 |
| | 6 | CUNI-MT | multiling | 42.1 | 45.9 | 67.4 |
| Unconstrain. | 1 | E-Translation | 2 | 56.5 | 33.2 | 75.4 |
| | | E-Translation | 1 | 56.0 | 33.5 | 75.2 |
| | 2 | PROMT | Transformer | 53.2 | 35.0 | 74.6 |
| | 3 | CdT-ASL | only-cdt-data | 51.4 | 37.0 | 72.9 |

**English–Italian**

| | Rank | Team | Description | BLEU [↑] | TER [↓] | BEER [↑] |
|---|---|---|---|---|---|---|
| Constrained | 1 | Lingua Custodia | 5lang-ov-ft-avg | 48.9 | 40.3 | 70.2 |
| | | PROMT | multilingual-model-round2-tuned-it | 48.3 | 39.5 | 70.4 |
| | 2 | Lingua Custodia | 5lang-ov | 48.0 | 40.9 | 69.8 |
| | 3 | E-Translation | 4bigTens | 47.0 | 41.7 | 69.7 |
| | | PROMT | multilingual-model-round2 | 46.8 | 40.6 | 69.1 |
| | 4 | E-Translation | 4bigTensFT | 46.7 | 42.2 | 68.9 |
| | 5 | Lingua Custodia | 1lang | 45.3 | 44.1 | 67.8 |
| | - | Baseline | Transformer+ | 43.5 | 44.0 | 67.9 |
| | - | Baseline | Transformer | 42.9 | 44.3 | 67.1 |
| | 6 | CUNI-MT | transfer | 38.6 | 48.1 | 64.6 |
| | 7 | CdT-ASL | only-round2-data | 37.9 | 51.9 | 62.6 |
| | 8 | PROMT | multilingual-model-round1 | 37.6 | 48.8 | 65.0 |
| | 9 | CUNI-MT | multiling | 35.2 | 53.1 | 62.5 |
| Unconstrained | 1 | E-Translation | 4bigTens | 50.1 | 39.0 | 71.0 |
| | 2 | E-Translation | 4bigTensnorm | 49.9 | 39.4 | 70.9 |
| | 3 | CdT-ASL | round2-data | 49.0 | 39.9 | 70.5 |
| | | PROMT | Transformer | 47.8 | 40.0 | 70.6 |
| | 4 | CdT-ASL | only-cdt-data | 45.2 | 43.3 | 68.8 |

**English–Modern Greek**

| | Rank | Team | Description | BLEU [↑] | TER [↓] | BEER [↑] |
|---|---|---|---|---|---|---|
| Constrained | 1 | PROMT | multilingual-model-round2-tuned-el | 45.1 | 42.3 | 67.8 |
| | 2 | Lingua Custodia | 7lang-ov-ft-avg | 44.7 | 43.8 | 67.2 |
| | 3 | Lingua Custodia | 7lang-ov | 44.2 | 44.1 | 67.0 |
| | 4 | Lingua Custodia | 7lang | 43.2 | 44.8 | 66.5 |
| | | PROMT | multilingual-model-round2 | 42.1 | 44.3 | 66.3 |
| | 5 | E-Translation | 1 | 41.7 | 46.2 | 65.5 |
| | 6 | Lingua Custodia | 1lang | 41.2 | 47.3 | 64.8 |
| | - | Baseline | Transformer+ | 39.8 | 46.9 | 64.7 |
| | - | Baseline | Transformer | 38.5 | 48.2 | 63.7 |
| | 7 | E-Translation | 2 | 34.9 | 53.2 | 60.8 |
| | | CUNI-MT | transfer | 34.9 | 51.6 | 61.3 |
| | 8 | CdT-ASL | only-round2-data | 32.9 | 56.6 | 59.0 |
| | | CUNI-MT | multiling | 32.4 | 56.1 | 59.5 |
| | | PROMT | multilingual-model-round1 | 31.4 | 55.2 | 59.5 |
| Unconst. | 1 | PROMT | Transformer | 44.4 | 44.0 | 67.2 |
| | 2 | E-Translation | 2 | 44.3 | 43.9 | 66.9 |
| | 3 | E-Translation | 1 | 43.1 | 44.7 | 66.3 |
| | 4 | CdT-ASL | only-cdt-data | 37.5 | 50.0 | 63.7 |

**English–Swedish**

| | Rank | Team | Description | BLEU [↑] | TER [↓] | BEER [↑] |
|---|---|---|---|---|---|---|
| Constrained | 1 | E-Translation | 4bigTens | 22.7 | 72.7 | 48.2 |
| | | Lingua Custodia | 5lang-ov-ft-avg | 22.0 | 71.7 | 49.2 |
| | | PROMT | multilingual-model-round2-tuned-sv | 21.8 | 69.3 | 49.7 |
| | | Lingua Custodia | 5lang-ov-r2-data | 21.8 | 71.5 | 49.4 |
| | 2 | PROMT | multilingual-model-round2 | 20.4 | 70.7 | 48.9 |
| | 3 | CdT-ASL | only-round2-data | 20.3 | 75.3 | 46.5 |
| | - | Baseline | Transformer+ | 19.5 | 72.2 | 48.1 |
| | 4 | Lingua Custodia | 7lang-ov-r1-data | 18.3 | 74.9 | 47.4 |
| | 5 | Lingua Custodia | 5lang-r1-data | 17.7 | 75.6 | 47.0 |
| | 6 | PROMT | multilingual-model-round1 | 17.2 | 75.3 | 46.7 |
| | 7 | Lingua Custodia | 1lang-r1-data | 16.7 | 78.9 | 45.4 |
| | | Baseline | Transformer | 15.3 | 77.5 | 44.4 |
| | 8 | CUNI-MT | multiling | 14.7 | 79.3 | 45.1 |
| | | CUNI-MT | transfer | 13.9 | 76.5 | 43.5 |
| Unconst. | 1 | E-Translation | 4bigTens | 23.3 | 70.2 | 50.0 |
| | 2 | CdT-ASL | only-cdt-data | 21.3 | 72.7 | 48.7 |
| | | PROMT | Transformer | 21.0 | 71.3 | 49.3 |

**English–Arabic**

| | Rank | Team | Description | BLEU [↑] | TER [↓] | BEER [↑] |
|---|---|---|---|---|---|---|
| Constrained | 1 | Lingua Custodia | 7lang-ov | 25.1 | 64.7 | 57.6 |
| | 2 | PROMT | multilingual-model-round2-tuned-ar | 22.9 | 62.9 | 56.5 |
| | 3 | Lingua Custodia | 7lang | 22.0 | 67.4 | 55.8 |
| | | PROMT | multilingual-model-round2 | 21.7 | 63.8 | 55.9 |
| | 4 | CUNI-MT | transfer | 19.1 | 68.7 | 52.9 |
| | | Lingua Custodia | 1lang | 19.1 | 73.8 | 53.0 |
| | | Baseline | Transformer | 18.8 | 69.3 | 52.3 |
| | 5 | CUNI-MT | multiling | 17.0 | 75.2 | 51.3 |
| | 6 | CdT-ASL | only-round2-data | 15.9 | 77.9 | 48.7 |
| Un. | 1 | PROMT | Transformer | 31.4 | 54.2 | 62.3 |

teams took part in this round. Among their approaches, the most successful ones were based on multilingual MT and transfer learning, using the Transformer architecture.

The second round addressed 7 different language pairs and was also divided into constrained and unconstrained scenarios, and engaged 5 different teams. Overall, a focus on the cleaning process of the data yielded great improvements. Most approaches were based on the Transformer and big Transformer architectures. Once more, multilingual models achieved great results, showing to be specially beneficial for languages with less resources. These results make sense due to the fact that Covid-19 corpora are needed to specialize models in the domain, even if they are from another language.

## Acknowledgments

The Covid-19 MLIA @ Eval initiative has received support from the European Commission, the European Language Resources Coordination (ELRC), European Language Resources Association (ELRA), the European Research Infrastructure for Language Resources and Technology (CLARIN), the CLEF Initiative and the Joint Research Centre (JRC). We gratefully acknowledge the translation team from Pangeanic for their help with the quality assessment.

## Author contribution

Khalid Choukri was one of the main organizers of the Covid-19 MLIA @ Eval initiative. Francisco Casacuberta, Miguel Domingo, Mercedes García-Martínez and Manuel Herranz organized the machine translation shared task. Alexandru Ceausu, Miltos Deligiannis, Guillaume Jacquet, Vassilis Papavassiliou, Stelios Piperidis, Prokopis Prokopidis, Dimitris Roussis and Marwa Hadj Salah were responsible of the data acquisition and engineering. Miguel Domingo and Mercedes García-Martínez wrote the main manuscript text. All authors reviewed the manuscript and contributed in generating the final version.

## References

[1] V. Papavassiliou, P. Prokopidis, G. Thurmair, A modular open-source focused crawler for mining monolingual and bilingual corpora from the web, in: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, 2013, pp. 43–51.

[2] D. Roussis, V. Papavassiliou, S. Sofianopoulos, P. Prokopidis, S. Piperidis, Constructing parallel corpora from covid-19 news using medisys metadata, 2022. Under review at LREC 2022.

[3] J. Linge, R. Steinberger, F. Fuart, S. Bucci, J. Belyaeva, M. Gemo, D. Al-Khudhairy, R. Yangarber, E. van der Goot, MediSys: medical information system. in advanced ICTs for disaster management and threat detection: Collaborative and distributed frameworks, IGI Global (2010) 131–142.

[4] G. Jacquet, M. Verile, Covid-19 news monitoring with medical information system (medisys), European Commission, Joint Research Centre (JRC). PID: http://data.europa.eu/89h/bd2f71e7-0551-4f57-8e82-fcfca8c1a462., 2020.

[5] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of the Association for Machine Translation in the Americas, 2006, pp. 223–231.

[6] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[7] M. Post, A call for clarity in reporting bleu scores, in: Proceedings of the Third Conference on Machine Translation, 2018, pp. 186–191.

[8] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, 2015, pp. 392–395.

[9] M. Stanojević, K. Sima'an, Fitting sentence level translation evaluation with many dense features, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 202–206.

[10] S. Riezler, J. T. Maxwell, On some pitfalls in automatic evaluation and significance testing for mt, in: Proceedings of the workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 57–64.

[11] L. Barrault, M. Biesialska, O. Bojar, M. R. Costa-jussà, C. Federmann, Y. Graham, R. Grundkiewicz, B. Haddow, M. Huck, E. Joanis, T. Kocmi, P. Koehn, C.-k. Lo, N. Ljubešić, C. Monz, M. Morishita, M. Nagata, T. Nakazawa, S. Pal, M. Post, M. Zampieri, Findings of the 2020 conference on machine translation (WMT20), in: Proceedings of the Fifth Conference on Machine Translation, 2020, pp. 1–55.

[12] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, volume 27, 2014, pp. 3104–3112.

[13] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2015. *arXiv:1409.0473*.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[15] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. M. Rush, OpenNMT: Open-Source Toolkit for Neural Machine Translation, in: Proceedings of the Association for Computational Linguistics: System Demonstration, 2017, pp. 67–72.

[16] F. A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: Continual prediction with LSTM, Neural computation 12 (2000) 2451–2471.

[17] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[19] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1715–1725.

[20] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, arXiv preprint arXiv:2001.08210 (2020).

[21] G. Lample, L. Denoyer, M. Ranzato, Unsupervised machine translation using monolingual corpora only, CoRR abs/1711.00043 (2017). URL: http://arxiv.org/abs/1711.00043. arXiv:1711.00043.

[22] T. Kocmi, O. Bojar, Trivial transfer learning for low-resource neural machine translation, in: Proceedings of the Third Conference on Machine Translation, 2018, pp. 244–252.

[23] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, A. Birch, Marian: Fast neural machine translation in C++, in: Proceedings of ACL 2018, System Demonstrations, 2018, pp. 116–121.

[24] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, T.-Y. Liu, Incorporating bert into neural machine translation, arXiv preprint arXiv:2002.06823 (2020).

[25] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, CoRR abs/1808.06226 (2018). URL: http://arxiv.org/abs/1808.06226. arXiv:1808.06226.

[26] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 2214–2218.

[27] A. Molchanov, PROMT systems for WMT 2019 shared translation task, in: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), 2019, pp. 302–307.

[28] C. Oravecz, K. Bontcheva, L. Tihanyi, D. Kolovratnik, B. Bhaskar, A. Lardilleux, S. Klocek, A. Eisele, etranslation's submissions to the wmt 2020 news translation task, in: Proceedings of the Fifth Conference on Machine Translation, 2020, pp. 254–261.