The Covid-19 MLIA @ Eval Initiative: Overview of the Machine Translation Task

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

[1] PRHLT Research Center - Universitat Politècnica de València, Spain
{fcn,midobal}@prhlt.upv.es
[2] European Commission
Alexandru.CEAUSU@ec.europa.eu
[3] Evaluations and Language resources Distribution Agency (ELDA), France
choukri@elda.org
[4] Athena Research Center, Greece
{mdel,vpapa,spip,prokopis}@athenarc.gr
[5] Pangeanic / B.I Europa - PangeaMT Technologies Division, Spain
{m.garcia,m.herranz}@pangeanic.com

Abstract. This report describes the Machine Translation task of the Covid-19 MLIA @ Eval initiative. The first round address 6 different language pairs (from English to German, French, Spanish, Italian, Modern Greek and Swedish) and was divided in two categories: one in which participants were limited to using only the provided corpora (constrained) and other in which it the use of external tools and data was allowed (unconstrained). 8 different teams took part in this round. The most successful approaches in both categories were based in multilingual machine translation and transfer learning. 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.

1 Introduction

In the current Covid-19 crisis, as in many other emergency situations, the general 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.

Covid-19 MLIA @ Eval Initiative, http://eval.covid19-mlia.eu/.

The goal of the Machine Translation (MT) task is to organize a community evaluation effort aimed at accelerating the creation of resources and tools for improving the generation of MT systems focused on Covid-19 related documents.

As the rest of the Covid-19 MLIA @ Eval initiative, we adopted an incremental and iterative evaluation methodology to enable the release of intermediate (but functional) resources and to progressively (iteration-after-iteration) move towards finally consolidated tools and resources. Thus, the task is divided in three rounds. At the end of each round, participants will write/update an incremental report explaining their system and highlighting which methods and data have been used.

2 MT Task Description

The goal of the MT task is to generate MT systems focused on Covid-19 related documents for different language pairs (which may differ for each round). Fig. 1 shows some examples of sentences from Covid-19 related documents.

30% of children and adults infected with measles can develop complications.

The MMR vaccine is safe and effective and has very few side effects.

The first dose is given between 10 and 18 months of age in European countries.

Note: The information contained in this factsheet is intended for the purpose of general information and should not be used as a substitute for the individual expertise and judgement of a healthcare professional.

Fig. 1. Examples of sentences from Covid-19 related documents.

Given a set of training data provided by the organizers for each language pair, participants have to train up to five different MT systems per language pair (see Section 3.1 for round 1). These systems are classified in two categories:

- **Constrained**: systems which have been trained exclusively 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 are allowed for this category.
- **Unconstrained**: systems which have been trained using data not provided by the organizers and/or any external resource not allowed in the constrained category.

Systems will be evaluated and compared according to the category to which they belong. It is mandatory that one of the submitted systems per language

pair belongs to the constrained category. Participants may take part in any or all of the language pairs. They will use their systems to translate a test set of unseen sentences in the source language. Evaluation will consist on assessing the translation quality of the submissions. Different criteria (e.g., automatic metrics) might be used on each round.

3 Round 1

3.1 Language Pairs

The first round of the Covid-19 MT task addressed the following language pairs:

- English–German.
- English–French.
- English–Spanish.
- English–Italian.
- English–Modern Greek.
- English–Swedish.

In all cases, the only translation direction was from English to the other language.

3.2 Data Generation

In the context of the first round of this initiative, we decided to generate 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 simulate a very quick response of the MT community in an emergency situation, like the current pandemic.

To this end, we first generated an updated version of the EMEA corpus[6] by harvesting the website of the European Medicines Agency[7], and applying 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[8], were processed in a similar manner and increased the volume of the "general" subset of the training data.

The first step of acquiring COVID-19-related data was the identification of multi bi-lingual websites with such content. With the aim of constructing data sets that could be publicly available, we targeted websites of national authorities and public health agencies (such a list is available at `https://`

[6] `https://elrc-share.eu/repository/browse/multilingual-corpus-made-out-of-pdf-documents-from-the-european-medicines-agency-emea-httpswwwemaeuropaeu-february-2020/3cf9da8e858511ea913100155d0267062d01c2d847c349628584d10293948de3/`.

[7] `https://www.ema.europa.eu/en`.

[8] `https://op.europa.eu/en/home`.

www.ecdc.europa.eu/en/COVID-19/national-sources), EU agencies and specific broadcast websites (e.g., https://voxeurop.eu/, https://globalvoices.org/, https://www.voltairenet.org/, etc.). In the next rounds we plan to also include relevant data from several international organizations and outcomes of broader crawls.

For acquiring domain-specific bilingual corpora, we used a recent version of ILSP-FC [15], 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[9] 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.

3.3 Corpora

The corpora was selected among the data generated in the previous section (see Section 3.2), splitting the documents 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 test documents 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. This procedure was conducted for each language pair. Table 1 contains the corpora statistics.

3.4 Evaluation

For this first run, evaluation was conducted automatically using two well-known MT metrics:

BiLingual Evaluation Understudy (BLEU) [16]: geometric average of the modified n-gram precision, multiplied by a brevity factor.

Character n-gram F-score (ChrF) [17]: character n-gram precision and recall arithmetically averaged over all character n-grams.

We used sacreBLEU [18] in order to ensure consistent scores. Additionally, we applied Approximate Randomization Testing (ART) [19]—with $10,000$ repetitions and using a p-value of 0.05—to determine whether two systems presented

[9] https://github.com/facebookresearch/LASER.

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

		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

statistically significance. The scripts used for conducting the automatic evaluation are publicly available together with some utilities which are useful for the shared task[10].

Among the two metrics, BLEU was selected as the main metric and, thus, it was used to rank the participants. Following the WMT criteria [2], 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 BLEU 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.

3.5 Baselines

For each language pair, we trained two different constrained systems to use as baselines: one based on recurrent neural networks (RNN) [1, 24] and another one based on the Transformer architecture [27]. All systems were built using `OpenNMT-py` [7].

Systems for the RNN baselines were trained using the standard parameters: long short-term memory units [3], with all model dimensions set to 512; Adam [6], with a fixed learning rate of 0.0002 and a batch size of 60; label smoothing of 0.1 [25]; beam search with a beam size of 6; and joint byte pair encoding (BPE) [21] applied to all corpora, using $32,000$ merge operations.

Similarly, systems for the Transformer baselines were trained using the standard parameters: 6 layers; Transformer [27], with all dimensions set to 512 except for the hidden transformer feed-forward (which was set to 2048); 8 heads of

[10] `https://github.com/midobal/covid19mlia-mt-task`.

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

Transformer self-attention; 2 batches of words in a sequence to run the generator on in parallel; a dropout of 0.1; Adam [6], 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 [25]; beam search with a beam size of 6; and joint byte pair encoding (BPE) [21] applied to all corpora, using $32,000$ merge operations.

3.6 Participants' Approaches

In this section, we present the different approaches submitted by the teams that took part in this round.

PROMT
PROMT's approaches consist in a multilingual model trained using Marian-NMT's [5] Transformer architecture.

For the constrained category, all data is concatenated using de-duplication to one single multilingual corpus to build a 8k SentencePiece[9] model for subword segmentation. In addition, a language-specific tag was added to the source side of the parallel sentence 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 appear less than ten times in the combined de-duplicated monolingual corpus from our vocabulary.

For the unconstrained category, all available data mainly from the OPUS[26] and statmt[12] with the addition of private data harvested from the Internet was added to the training data. A special BPE implementation[13] developed by the team was applied instead of SentencePiece but the author used SentencePiece in the constrained option as it seems to work better in low-resource settings. The size of the BPE models and vocabularies varies from 8k to 16k and shared vocabulary is not used (separate BPE models are trained) for the English–Greek pair as the two languages have different alphabets.

The participant submitted systems for all the language pairs and constrained and unconstrained categories. PROMT's systems rank first in all but the English–German unconstrained category. The team plans to tune their baseline systems for the second round.

CUNI-MT
This team submitted 3 different approaches to the constrained category: 1) standard Neural Machine Translation (NMT) training with back-translation; 2) transfer learning; and 3) multilingual training.

1. Their standard NMT approach relies on one bi-directional model (sharing the encoder and decoder for both translation directions) which constantly switches between training and inference mode to produce batches of synthetic sentence pairs, learning from both authentic and synthetic training samples

[12] `http://www.statmt.org/`.

using online back-translation[10]. The models are trained on BPE units[21] with a vocabulary of 30k items.

2. Their second approach consists of the transfer learning approach proposed by Kocmi and Bojar[8] (one of the participants), who fine-tuned a low-resource child model from a pre-trained high-resource parent model for a different language pair. The subword vocabulary generated from the child and parent language pair corpora is shared.
   The training procedure consists of first training an NMT model on the parent parallel corpus until it converges. Then, they replace the training data with the child corpus. They experimented repeating this procedure several times with the child becoming the parent for either a completely new language (e.g., German $\rightarrow$ English $\rightarrow$ Spanish $\rightarrow \cdots$) or for the original parent (e.g., German $\rightarrow$ English $\rightarrow$ German $\rightarrow \cdots$). When adding a new language, the joint BPE vocabulary has to be modified by replacing the original parent vocabulary entries with the new children.
3. Their multilingual approach consists of a model trained to translate from English to French, Italian and Spanish (due to language similarities). During inference, the corresponding embedding of the target language is selected. The BPE vocabulary was extracted from the concatenation of all four corpora, using only unique English sentences to reach a comparable corpus size.

The training was performed using the XLM[13] toolkit and the vocabulary size was set to 30k. CUNI-MT's systems ranked first for constrained English–German and constrained English–Swedish.

CUNI-MTIR
This team submitted systems for English into French, German, Swedish and Spanish in both constrained and unconstrained settings. Transformer architecture from MarianNMT [5] toolkit was used in order to train the models.

For unconstrained systems, they used UFAL Medical Corpus[14] for training data and then fine-tuned the models with the constrained data.

All the data is tokenized using Khresmoi[15]'s tokenizer and, then, encoded using BPE with 32K merges.

Lingua Custodia (LC)
LC submissions consisted of a multilingual model able to translate from English to French, German, Spanish, Italian and Swedish; and individual translation models for English to German and French language pairs.

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

[13] `https://github.com/facebookresearch/XLM`.
[14] `http://ufal.mff.cuni.cz/ufal_medical_corpus`.
[15] `http://www.khresmoi.eu/assets/Deliverables/WP4/KhresmoiD412.pdf`.

the target language. The English to German multilingual model achieved much higher score than the English to German single model. This improvement is not shown in the English to French model.

The LC system ranked first for constrained English[5]Swedish. For next rounds, they plan to use transfer learning from massive language models.

LIMSI

LIMSI submitted systems to English to French constrained and unconstrained categories. 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[16] general data and fine-tuned on the shared task's corpus; 3) same as 2) but adding BERT[28]; and 4) a system only trained with constrained data but computing the BPE codes using all the external in-domain corpus.

The systems are trained using Transformer architecture from *fairseq*[17] (Facebook's seq-2-seq library).

TARJAMA-AI

Tarjama-AI submitted a single system for English to Spanish, German, Italian, French and Swedish constrained category. This system consisted of a model trained with all the language pairs data adding a special token for the non-target languages. Additionally, they oversampled the corpus of the desired target language (i.e., the English–Spanish corpus for training the constrained English–Spanish, etc).

E-Translation

E-Translation submitted systems for English–German and English–French language pairs. They used Transformer models from MarianNMT toolkit [5].

For constrained English–German, they used transfer learning and 12K size vocabulary created using SentencePiece. For the unconstrained category, they submitted their WMT system and a new version of that system, fine-tuned with the constrained data.

They also participated in constrained English–French with two systems described as *small* and *big*, and in unconstrained English–French with three systems described as *gen*, *phwt* and *eufl*. Their system ranked first for unconstrained English–German.

ACCENTURE

This participants' report is missing but they noted in their system's description that they used multilingual BART[12].

[16] `http://www.statmt.org/wmt14/translation-task.html`.
[17] `https://fairseq.readthedocs.io/en/latest/models.html`.

3.7 Results

In this section, we present the results from the first round. Following the WMT criteria [2], we grouped systems together into clusters according to which systems significantly outperformed all others in lower ranking clusters, according to ART (see Section 4.4).

Overall, *multilingual* and *transfer learning* approaches yielded the best results for all languages pairs in the constrained category. 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 category, it had less participation than the constrained one. With the exemption of English–German—in which *ETRANSLATION* approaches based on their WMT system [14]yielded better results—*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.

English–German

Table 2 presents the results for the English–German language pair. 12 different systems from 6 participants were submitted to the constrained category, and 4 different systems from 3 participants were submitted to the unconstrained.

The best results for the constrained category were achieved by three of *CUNI-MT*'s systems and *PROMT* (who submitted a single system for this language pair). Their approaches were based on transfer learning, standard NMT with back-translation and multilingual NMT. The next best results were achieved by *ETRANSLATION*'s system, which used a transfer learning approach. Then, we have another of *CUNI-MT*'s transfer learning approaches and *LC*'s multilingual approach. On fourth position we have our Transformer baseline and *LC*'s Transformer approach. On fifth and sixth positions are *TARJAMA-AI*'s approaches based on tagged back-translation and combining all language pairs (adding a special token to all sentences except the ones from English–German). Next we have *CUNI-MTIR*'s Transformer approach. Finally, our RNN baseline

Table 2. Results of the English–German language pair, 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.

	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	ETRANSLATION	basetr	30.4	0.593
	3	CUNI-MT	transfer2	29.8	0.584
		LC	multilingual	29.5	0.584
	4	Baseline	transformer	28.1	0.573
		LC	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	ETRANSLATION	wmtfinetune	44.4	0.686
	2	ETRANSLATION	wmt	44.1	0.683
	3	PROMT	transformer	41.2	0.666
	4	CUNI-MTIR	r1	20.0	0.499

and *TARJAMA-AI*'s NMT approach (they did not specify the architecture they used to train their system) placed last.

For the unconstrained category, the best results were achieved by *ETRANSLATION*'s WMT system fine-tuned with the in-domain data[18]. Second position was for *ETRANSLATION*'s WMT system. At third place, we have *PROMT*'s multilingual system. Finally, we have *CUNI-MTIR*'s Transformer approach. It is worth noting how *ETRANSLATION*'s WMT system—which has been trained using exclusively out-of-domain data—achieved around 13 BLEU points of improvement over the best constrained system. We discussed this phenomenon during the virtual meeting and came to the conclusion that, despite that we are working in a very specific domain (Covid-19 related documents), the subgenre of this domain[19] is more closely related to the news domain of WMT than expected.

[18] We are waiting for their report to know more details about their approach.

[19] Note that the target audience of Covid-19 MLIA is the general public. Thus, while documents are indeed Covid-19 related, they lean towards information aimed at the citizens instead of scientific or medical documents.

English–French

Table 3 presents the results for the English–French language pair. 12 different systems from 8 participants were submitted to the constrained category, and 8 different systems from 4 participants to the unconstrained one.

Table 3. Results of the English–French language pair, 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.

	Rank	Team	Description	BLEU [↑]	chrF [↑]
Constrained	1	PROMT	multilingual	49.6	0.711
	2	ETRANSLATION	small	49.1	0.707
		LC	multilingual	49.0	0.705
		LC	transformer	48.9	0.703
		CUNI-MT	base	48.4	0.703
		CUNI-MT	multiling	48.0	0.700
		ETRANSLATION	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	ETRANSLATION	gen	52.9	0.742
	3	LIMSI	indom	51.2	0.721
	4	ETRANSLATION	phwt	50.1	0.724
		LIMSI	trans	49.3	0.710
		LIMSI	bert	49.3	0.703
		LIMSI	mlia	48.5	0.705
	5	ETRANSLATION	eufl	47.9	0.712
	6	CUNI-MTIR	r1	33.0	0.590

PROMT's multilingual approach yielded the best results for the constrained category. Second on the ranking are *ETRANSLATION*'s *small* and *big* approaches[20], *LC*'s multilingual and Transformer approaches, *CUNI-MT*'s back-translation, multilingual and transfer learning approaches and our Transformer baseline. Finally—placing each one on a different rank—we have *LIMSI*'s trans-

[20] They have yet to provide a description of their approaches.

fer learning approach, *CUNI-MTIR*'s Transformer approach, our RNN baseline, *TARJAMA-AI*'s NMT approach and *ACCENTURE*'s multilingual bart approach.

The best unconstrained results were achieved by *PROMT*'s multilingual system. Following is *ETRANSLATION*'s *gen* approach. Then, we have *LIMSI*'s approach based on Transformer using in-domain corpora. On fourth place, we have *ETRANSLATION*'s *phwt* approach and *LIMSI*'s approaches based on using out-of-domain corpora—with and without the use of BERT—fine-tuned with the provided data set, and their approach based on training exclusively with the provided data set, but training BPE using additional in-domain corpora. Then, we have *ETRANSLATION*'s *eufl* approach. Finally, we have *CUNI-MTIR*'s Transformer approach.

English–Spanish

Table 4 presents the results for the English–Spanish language pair. 9 different systems from 6 participants were submitted to the constrained category, and only 2 different systems from 2 participants for the unconstrained one.

Table 4. Results of the English–Spanish language pair, 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.

	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
		LC	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

PROMT's multilingual approach yielded the best constrained results. Second in the rank we have *CUNI-MT*'s transfer learning, multilingual and back-translation approaches, *LC*'s multilingual approach and our Transformer baseline. Following up is our RNN baseline. Finally, on third, fourth and fifth po-

sitions we have *CUNI-MTIR*'s Transformer approach, *TARJAMA-AI*'s NMT approach and *ACCENTURE*'s multilingual bart approach (respectively).

For the unconstrained category, the best results were achieved by *PROMT*'s multilingual system, followed by *CUNI-MTIR*'s Transformer approach.

English–Italian

Table 5 presents the results for constrained English–Italian language pair. 5 different systems from 4 participants were submitted for the constrained category, and a single system to the unconstrained one.

Table 5. Results of the English–Italian language pair, 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.

	Rank	**Team**	**Description**	**BLEU** [↑]	**chrF** [↑]
Constrained	1	PROMT	multilingual	29.6	0.585
	2	LC	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

PROMT's multilingual approach yielded the best constrained results. Next, sharing position two, we have *LC*'s multilingual approach and *CUNI-MT*'s transfer learning and multilingual approaches. After that, we have our Transformer baseline. On third position we have *TARJAMA-AI*'s NMT approach. Finally, we have our RNN baseline.

Regarding the unconstrained category, only *PROMT*'s multilingual approach was submitted.

English–Modern Greek

Table 6 presents the results for English–Modern Greek language pair. 3 different systems from 2 participants were submitted for the constrained category, and a single system was submitted to the unconstrained one. Thus, this was the language pair with less participation. According to participant's reports, this was mostly due to Modern Greek using a different alphabet.

Once more, *PROMT*'s multilingual approach yielded the best constrained results. Second, we have *CUNI-MT*'s transfer learning approach. On the third position we have *CUNI-MT*'s back-translation approach, which shares cluster with our Transformer baseline. Finally, we have our RNN baseline.

Table 6. Results of the English–Modern Greek language pair, 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.

	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

Only *PROMT*'s multilingual approach was submitted to the unconstrained category.

English–Swedish

Table 7 presents the results for the English–Swedish language pair. 7 different systems from 5 participants were submitted to the constrained category, and two system from two participants to the unconstrained one.

Table 7. Results of the English–Swedish language pair, 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.

	Rank	Team	Description	BLEU [↑]	chrF [↑]
Constrained	1	PROMT	multilingual	30.7	0.595
		LC	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

The best results for the constrained category were yielded by *PROMT*'s multilingual approach, *LC*'s multilingual approach and one of *CUNI-MT*'s transfer

learning approaches. The other *CUNI-MT*'s transfer learning approach placed second on the ranking. Next we have our Transformer baseline. Third position is taken by *CUNI-MT*'s back-translation approach. Next we have *CUNI-MTIR*'s Transformer approach. Following up is our RNN baseline. Finally, on fifth position we have *TARJAMA-AI*'s NMT approach.

Regarding the unconstrained category, the best results were achieved by *PROMT*'s multilingual system, followed by *CUNI-MTIR*'s Transformer approach.

3.8 Human Quality Assessment

Taking into account that the corpora used for this round was obtained from crawling (see Section 3.2), it is important to assess the quality of the reference sets. To do so, we selected a subset of the Spanish corpora and post-edited it with the help of a team of professional translators. This subset consisted in 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*.

As a first step towards assessing the quality of the reference sets, we compared the reference and its post-edited version using human Translation Error Rate (hTER) [22]. 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. Table 8 shows the results of this evaluation. We obtained a fairly low TER value (18.8), which indicates that the translation quality of the reference is generally good and, thus, is coherent with the translators' opinion.

Table 8. Evaluation of the translation quality of the Spanish reference set.

	hTER
reference	18.8

As a second step, we re-evaluated participant's translations (the corresponding subset only) using both the reference and its post-edited version. Table 9 present the results of the evaluation. In all cases, both metrics show fairly similar results—with a preference towards the reference, which is to be expected since its style is more similar to the training data. Thus, we can conclude that the quality of the reference sets is proficient enough to be used in an automatic evaluation and that the results obtained in the previous section (see Section 4.7) are significant.

Table 9. Results of evaluating a subset of the Spanish test using either the reference or its post-edited version.

Team	Description	Reference		Post-edition	
		BLEU [↑]	**chrF** [↑]	**BLEU** [↑]	**chrF** [↑]
PROMT	multilingual	45.1	0.682	43.9	0.672
CUNI-MT	transfer1	46.2	0.686	43.8	0.672
CUNI-MT	transfer2	46.0	0.686	43.4	0.671
LC	multilingual	45.8	0.684	43.5	0.669
Baseline	transformer	45.4	0.682	43.9	0.670
CUNI-MT	multiling	44.7	0.677	43.0	0.664
CUNI-MT	base	45.0	0.675	42.4	0.660
Baseline	RNN	34.6	0.603	32.3	0.589
CUNI-MTIR	r1	31.4	0.583	30.8	0.578
TARJAMA-AI	base	29.2	0.583	26.9	0.569
ACCENTURE	mbart	16.7	0.466	16.0	0.460

3.9 Conclusions

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

8 different 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. PROMT's approach yielded the best results for all language pairs in both categories except for unconstrained English–German and constrained English–German—in which they shared the first position with CUNI-MT.

In general, there were not big differences between systems of consecutive ranks (according to both metrics). We provided two different baselines which worked well as delimiters: more sophisticated approaches 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).

4 Round 2

4.1 Language Pairs

The second round of the Covid-19 MT task addressed the following language pairs:

– English–German.

– English–French.
– English–Spanish.
– English–Italian.
– English–Modern Greek.
– English–Swedish.
– English–Arabic.

In all cases, the only translation direction was from English to the other language.

4.2 Data Generation

In the context of the second round of this initiative, we decided to exploit the outcomes of an available infrastructure, namely MediSys (Medical Information System), with the purpose of constructing parallel corpora beneficial for MT [20]. Similarly to the first round's approach, it could be seen as an application of simulating 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 [11]. Focusing on the current pandemic, a dataset of metadata which concerns COVID-19 related news was made publicly available in RSS/XML format, and corresponds to millions of news articles [4]. It is worth mentioning that the dataset is 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 are COVID-19 related monolingual corpora and could be 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 deduplicates, etc).

In addition, we re-crawled several websites of national authorities and public health agencies in order to enrich the data that have been collected during the first round (see section 3.2).

4.3 Corpora

Given the data generated in the previous section (see Section 3.2), 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. 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 reasons, 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 contained low-quality segments (remember that the data has 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 representation (%) 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 established 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 quality.
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 10 contains the corpora statistics.

Table 10. Corpora statistics. $|S|$ stands for number of sentences, $|T|$ for number of tokens and $|V|$ for size of the vocabulary. M denotes millions and K thousands.

		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

4.4 Evaluation

For this second run, evaluation was conducted automatically using three well-known MT metrics:

BiLingual Evaluation Understudy (BLEU) [16]: geometric average of the modified n-gram precision, multiplied by a brevity factor.

Translation Edit Rate (TER) [22]: 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) [23]: a sentence level metric that incorporates a large number of features combined in a linear model.

We used `sacreBLEU` [18] in order to ensure consistent scores. Additionally, we applied Approximate Randomization Testing (ART) [19]—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 are useful for the shared task[21].

Following the WMT criteria [2], 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[22]. Otherwise, we added it into a new cluster. This way, systems from one cluster significantly outperformed all others in lower ranking clusters.

4.5 Baselines

For each language pair, we trained two different constrained systems to use as baselines: one using only the second round data and other using also the first round data (except for En–Ar, for which only round 2 data is available).

Systems were built using `OpenNMT-py` [7] and are based on the Transformer architecture [27]. They were trained using the standard parameters: 6 layers; Transformer [27], 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 [6], 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 [25]; beam search with a beam size of 6; and joint byte pair encoding (BPE) [21] applied to all corpora, using $32,000$ merge operations.

4.6 Participants' Approaches

In this section, we present the different approaches submitted by the teams that took part in the second round.

LC
Lingua Custodia team participated in the constrained translation task. The preprocessing 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

[21] `https://github.com/midobal/covid19mlia-mt-task`.
[22] Considering that, at the start of this process, there is an initial cluster containing the first system.

adding a tag with the casing information. They finally append the language token to each source sentence in the preprocessing 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.

They tested different settings trying different vocabulary sizes and obtained better results for 40K and 50K. Another experiment resulted in bettter results using preprocessing. They also compared bilingual models with multilingual models using the 5 languages excluding Arabic and Greek due to the different scripting or a multilingual models with the 7 languages. Languages with most amount of data (Spanish, French and German show show better results using bilingual models and their score is reduced using multilingual and even more when oversampling less resource languages. On the contrary, Italian and Swedish benefits from multilingual models and even more using oversampling. When using 7 languages multilingual model benefit Greek and Arabic and again more with oversampling.

Their final results show that Spanish bilingual model obtained the first position in the ranking and also German doing finetuning with the 5 languages multilingual models, French obtained 3rd position. Italian and Swedish obtained best results with oversampling and finetuning 5 languages multilingual model with 1st position in Italian and 2nd position in the ranking in Swedish. Lastly, Arabic and Greek obtained better results with the 7 languages multilingual model obtaining 1st position for Arabic and 2nd for Greek.

ETRANSLATION

ETranslation participated in contrained and uncontrained options for 6 language pairs (all but Arabic). They performed 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 postprocessing, only sanity checking. They experimented with standard Transformer and big Transformer in MarianNMT for contrained models. In their first submission of unconstrained option, models were trained adding TAUS Corona Crisis Corpora the OPUS EMEA Corpus and a health related subset of the Euramis data set. For German, the second submission was the one used in first round where they used WMT 2021 news task data and fine tuned with in-domain data. Surprisingly, the second submission obtained the highest score and it is only one point better than the model trained without in-domain data. This can be due to the fact that in-domain data comes from sources of information about covid-19 and it can be a closed domain to news and WMT dataset is much bigger than the rest of the datasets used in the first submission. The best architecture was a 4 model big transformer ensemble. Spanish systems were more competitive than Greek ones with the same setting. For Italian they used part of the traning dataset extracted using keywords for fine-tuning, this only improved dev datatest results. For French uncontrained submission, they used the stock eTranslation general

engine trained with Euramis and OPUS data. They also noticed that a postprocess to nomalize puntuation improved by 7 BLEU points in uncontrained option. E-translation ranked the first in all the language pairs that they participated but Greek in unconstrained mode and for French and Swedish for contrained mode.

CdT-ASL

CdT-ASL team developed NICE which integrates neural machine translation custom engines for confidencial adapted translations. They submitted constrained and unconstrained systems, they added generic and public health domains CdT own data for uncontrained systems. They apply cleaning processes to prepare the data for training with big transformer using OpenNMT-tf.

PROMT

PROMPT trained a transformer multilingual model with a single encoder and a single decoder with Marian toolkit. For round 2, PROMPT did a fine-tune for the language pairs improving 1-2 additional BLEU points in contrained mode. Unconstrained mode remains as round 1. They ranked top in English-Greek and show competitive results in other language pairs.

CUNI-MT

CUNI MT team trained their multilingual models using Transformer in MarianNMT toolkit. They trained jointly on all languages. The results for contrained mode showed better results for transfer learning models. They concluded that when pretraining a model on a different language pair better results are obtained when the corpus size is big and the transfer works also for completely unrelated languages.

4.7 Results

In this section, we present the results from the second round. Following the WMT criteria [2], we grouped systems together into clusters according to which systems significantly outperformed all others in lower ranking clusters, according to ART (see Section 4.4).

Results are grouped according to BLEU, TER and BEER.

English–German
Table 11 presents the results for the English–German language pair according to BLEU.

English–French
Table 14 presents the results for the English–French language pair according to BLEU.

Table 11. Results of the English–German language pair, 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. ⋆ denotes teams which due to not fulfilling the requirement of submitting at least one constrained system.

	Rank	Team	Description	BLEU [↑]
Constrained	1	LC	5lang-ft-avg	40.3
		ETRANSLATION	ensembleFT	39.9
		LC	5lang-ft	39.8
		ETRANSLATION	ensemble	39.7
		LC	1lang	39.7
	2	PROMT	multilingual-model-round2-tuned-de	39.6
	3	LC	7lang	38.6
	4	PROMT	multilingual-model-round2	39.6
	-	Baseline	Transformer	34.9
		Baseline	Transformer+	34.8
	5	CUNI-MT	transfer	31.8
	6	PROMT	multilingual-model-round1	28.7
	7	CUNI-MT	transfer2	27.5
	8	CUNI-MT	multiling	27.0
Unconstrained	1	ETRANSLATION	wmtFT	45.7
	2	PROMT	Transformer	40.4
	3	ETRANSLATION	singlebigTr	40.0
	4	ETRANSLATION	eTstandardengine	35.4
	-	CdT-ASL⋆	only-cdt-data	34.9

English–Spanish
Table 17 presents the results for the English–Spanish language pair according to BLEU.

English–Italian
Table 20 presents the results for the English–Italian language pair according to BLEU.

English–Modern Greek
Table 20 presents the results for the English–Modern Greek language pair according to BLEU.

Table 12. Results of the English–German language pair, divided by categories. Systems are ranked according to TER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally. ⋆ denotes teams which due to not fulfilling the requirement of submitting at least one constrained system.

	Rank	Team	Description	TER [↓]
Constrained	1	PROMT	multilingual-model-round2-tuned-de	47.7
	2	ETRANSLATION	ensembleFT	48.2
		ETRANSLATION	ensemble	48.4
		LC	5lang-ft-avg	48.4
		LC	5lang-ft	48.9
	3	PROMT	multilingual-model-round2	49.6
		LC	7lang	50.0
		LC	1lang	50.1
	-	Baseline	Transformer	51.7
		Baseline	Transformer+	51.8
	4	CUNI-MT	transfer	54.6
	5	PROMT	multilingual-model-round1	57.7
	6	CUNI-MT	transfer2	60.4
	7	CUNI-MT	multiling	60.9
Unconstrained	1	ETRANSLATION	wmtFT	43.0
	2	PROMT	Transformer	46.9
	3	ETRANSLATION	singlebigTr	48.4
	4	ETRANSLATION	eTstandardengine	52.7
	-	CdT-ASL⋆	only-cdt-data	53.1

English–Swedish

Table 20 presents the results for the English–Swedish language pair according to BLEU.

English–Arabic

Table 20 presents the results for the English–Arabic language pair according to BLEU.

Table 13. Results of the English–German language pair, divided by categories. Systems are ranked according to BEER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally. ⋆ denotes teams which due to not fulfilling the requirement of submitting at least one constrained system.

	Rank	Team	Description	BEER [↑]
Constrained	1	LC	5lang-ft-avg	66.8
		ETRANSLATION	ensembleFT	66.8
		PROMT	multilingual-model-round2-tuned-de	66.8
		ETRANSLATION	ensemble	66.6
		LC	5lang-ft	66.5
	2	LC	1lang	65.9
		LC	7lang	65.8
		PROMT	multilingual-model-round2	65.7
	-	Baseline	Transformer	63.9
		Baseline	Transformer+	63.7
	3	CUNI-MT	transfer	61.8
	4	PROMT	multilingual-model-round1	60.8
	5	CUNI-MT	transfer2	59.8
	6	CUNI-MT	multiling	59.2
Unconstrained	1	ETRANSLATION	wmtFT	70.4
	2	PROMT	Transformer	67.9
	3	ETRANSLATION	singlebigTr	66.9
	4	ETRANSLATION	eTstandardengine	64.6
	-	CdT-ASL⋆	only-cdt-data	64.3

Table 14. Results of the English–French language pair, 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.

	Rank	Team	Description	BLEU [↑]
Constrained	1	ETRANSLATION	2	58.3
		ETRANSLATION	1	57.9
		LC	1lang	57.2
		PROMT	multilingual-model-round2-tuned-fr	57.1
	2	CdT-ASL	only-round2-data	56.9
	3	LC	7lang	55.8
		PROMT	multilingual-model-round2	55.4
	-	Baseline	Transformer	54.4
	-	Baseline	Transformer+	53.7
	4	PROMT	multilingual-model-round1	45.4
	5	CUNI-MT	multiling	44.1
Unconstrained	1	PROMT	Transformer	57.1
		ETRANSLATION	generaldenorm	56.9
	2	ETRANSLATION	general	49.9
		CdT-ASL	only-cdt-data	49.7
	3	ETRANSLATION	formal	43.5

Table 15. Results of the English–French language pair, divided by categories. Systems are ranked according to TER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	TER [↓]
Constrained	1	ETRANSLATION	2	33.8
		ETRANSLATION	1	34.0
		PROMT	multilingual-model-round2-tuned-fr	34.1
	2	CdT-ASL	only-round2-data	34.6
		LC	1lang	34.9
		PROMT	multilingual-model-round2	35.2
		LC	7lang	35.7
		Baseline	Transformer	35.9
	-	Baseline	Transformer+	36.7
	3	PROMT	multilingual-model-round1	43.1
	4	CUNI-MT	multiling	45.6
Unconstrained	1	PROMT	Transformer	34.5
		ETRANSLATION	generaldenorm	34.8
	2	ETRANSLATION	general	38.8
	3	CdT-ASL	only-cdt-data	40.0
	4	ETRANSLATION	formal	44.6

Table 16. Results of the English–French language pair, divided by categories. Systems are ranked according to BEER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	BEER [↑]
Constrained	1	ETRANSLATION	2	75.1
		ETRANSLATION	1	75.0
		PROMT	multilingual-model-round2-tuned-fr	74.8
	2	LC	1lang	74.5
		CdT-ASL	only-round2-data	74.5
	3	PROMT	multilingual-model-round2	74.0
		LC	7lang	73.9
	-	Baseline	Transformer	73.4
	-	Baseline	Transformer+	73.0
	4	PROMT	multilingual-model-round1	68.9
	5	CUNI-MT	multiling	67.6
Unconstrained	1	PROMT	Transformer	74.8
	2	ETRANSLATION	generaldenorm	74.5
	3	ETRANSLATION	general	72.0
	4	CdT-ASL	only-cdt-data	71.3
	5	ETRANSLATION	formal	68.0

Table 17. Results of the English–Spanish language pair, 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.

	Rank	Team	Description	BLEU [↑]
Constrained	1	LC	1lang-avg	56.6
		ETRANSLATION	2	56.1
		ETRANSLATION	1	56.1
		LC	5lang-ft-avg	56.0
	2	CdT-ASL	only-round2-data	55.4
		LC	7lang	55.3
		PROMT	multilingual-model-round2-tuned-es	54.9
	3	PROMT	multilingual-model-round2	53.8
	-	Baseline	Transformer	53.3
	-	Baseline	Transformer+	51.8
	4	CUNI-MT	transfer	48.4
	5	PROMT	multilingual-model-round1	45.1
	6	CUNI-MT	multiling	42.1
Unconstrain.	1	ETRANSLATION	2	56.5
		ETRANSLATION	1	56.0
	2	PROMT	Transformer	53.2
	3	CdT-ASL	only-cdt-data	51.4

Table 18. Results of the English–Spanish language pair, divided by categories. Systems are ranked according to TER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	TER [↓]
Constrained	1	ETRANSLATION	2	33.5
		ETRANSLATION	1	33.5
		LC	1lang-avg	33.7
		LC	5lang-ft-avg	33.8
		PROMT	multilingual-model-round2-tuned-es	33.9
		CdT-ASL	only-round2-data	34.1
		LC	7lang	34.4
		PROMT	multilingual-model-round2	34.5
	-	Baseline	Transformer	35.2
	-	Baseline	Transformer+	36.1
	2	CUNI-MT	transfer	39.3
	3	PROMT	multilingual-model-round1	41.2
	4	CUNI-MT	multiling	45.9
Unconstrain.	1	ETRANSLATION	2	33.2
		ETRANSLATION	1	33.5
	2	PROMT	Transformer	35.0
	3	CdT-ASL	only-cdt-data	37.0

Table 19. Results of the English–Spanish language pair, divided by categories. Systems are ranked according to BEER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	BEER [↑]
Constrained	1	LC	1lang-avg	75.2
		ETRANSLATION	2	75.2
		ETRANSLATION	1	75.2
		LC	5lang-ft-avg	75.1
	2	PROMT	multilingual-model-round2-tuned-es	74.9
		LC	7lang	74.8
		CdT-ASL	only-round2-data	74.6
	3	PROMT	multilingual-model-round2	74.3
	-	Baseline	Transformer	74.0
	-	Baseline	Transformer+	73.3
	4	CUNI-MT	transfer	71.2
	5	PROMT	multilingual-model-round1	69.9
	6	CUNI-MT	multiling	67.4
Unconstrain.	1	ETRANSLATION	2	75.4
		ETRANSLATION	1	75.2
	2	PROMT	Transformer	74.6
	3	CdT-ASL	only-cdt-data	72.9

Table 20. Results of the English–Italian language pair, 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.

	Rank	Team	Description	BLEU [↑]
Constrained	1	LC	5lang-ov-ft-avg	48.9
		PROMT	multilingual-model-round2-tuned-it	48.3
	2	LC	5lang-ov	48.0
	3	ETRANSLATION	4bigTens	47.0
		PROMT	multilingual-model-round2	46.8
	4	ETRANSLATION	4bigTensFT	46.7
	5	LC	1lang	45.3
	-	Baseline	Transformer+	43.5
	-	Baseline	Transformer	42.9
	6	CUNI-MT	transfer	38.6
	7	CdT-ASL	only-round2-data	37.9
	8	PROMT	multilingual-model-round1	37.6
	9	CUNI-MT	multiling	35.2
Unconstrained	1	ETRANSLATION	4bigTens	50.1
	2	ETRANSLATION	4bigTensnorm	49.9
	3	CdT-ASL	round2-data	49.0
		PROMT	Transformer	47.8
	4	CdT-ASL	only-cdt-data	45.2

Table 21. Results of the English–Italian language pair, divided by categories. Systems are ranked according to TER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	TER [↓]
Constrained	1	PROMT	multilingual-model-round2-tuned-it	39.5
	2	LC	5lang-ov-ft-avg	40.3
		PROMT	multilingual-model-round2	40.6
		LC	5lang-ov	40.9
	3	ETRANSLATION	4bigTens	41.7
	4	ETRANSLATION	4bigTensFT	42.2
	5	Baseline	Transformer+	44.0
		LC	1lang	44.1
		Baseline	Transformer	44.3
	6	CUNI-MT	transfer	48.1
		PROMT	multilingual-model-round1	48.8
	7	CdT-ASL	only-round2-data	51.9
	8	CUNI-MT	multiling	53.1
Unconstrained	1	ETRANSLATION	4bigTens	39.0
	2	ETRANSLATION	4bigTensnorm	39.4
	3	CdT-ASL	round2-data	39.9
		PROMT	Transformer	40.0
	4	CdT-ASL	only-cdt-data	43.3

Table 22. Results of the English–Italian language pair, divided by categories. Systems are ranked according to BEER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	BEER [↑]
Constrained	1	PROMT	multilingual-model-round2-tuned-it	70.4
		LC	5lang-ov-ft-avg	70.2
	2	LC	5lang-ov	69.8
		ETRANSLATION	4bigTens	69.7
	3	PROMT	multilingual-model-round2	69.1
		ETRANSLATION	4bigTensFT	68.9
	4	Baseline	Transformer+	67.9
		LC	1lang	67.8
	-	Baseline	Transformer	67.1
	5	PROMT	multilingual-model-round1	65.0
		CUNI-MT	transfer	64.6
	6	CdT-ASL	only-round2-data	62.6
		CUNI-MT	multiling	62.5
Unconstrained	1	ETRANSLATION	4bigTens	71.0
	2	ETRANSLATION	4bigTensnorm	70.9
		PROMT	Transformer	70.6
		CdT-ASL	round2-data	70.5
	3	CdT-ASL	only-cdt-data	68.8

Table 23. Results of the English–Modern Greek language pair, 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.

	Rank	Team	Description	BLEU [↑]
Constrained	1	PROMT	multilingual-model-round2-tuned-el	45.1
	2	LC	7lang-ov-ft-avg	44.7
	3	LC	7lang-ov	44.2
	4	LC	7lang	43.2
		PROMT	multilingual-model-round2	42.1
	5	ETRANSLATION	1	41.7
	6	LC	1lang	41.2
	-	Baseline	Transformer+	39.8
	-	Baseline	Transformer	38.5
	7	ETRANSLATION	2	34.9
		CUNI-MT	transfer	34.9
	8	CdT-ASL	only-round2-data	32.9
		CUNI-MT	multiling	32.4
		PROMT	multilingual-model-round1	31.4
Unconst.	1	PROMT	Transformer	44.4
	2	ETRANSLATION	2	44.3
	3	ETRANSLATION	1	43.1
	4	CdT-ASL	only-cdt-data	37.5

Table 24. Results of the English–Modern Greek language pair, divided by categories. Systems are ranked according to TER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	TER [↓]
Constrained	1	PROMT	multilingual-model-round2-tuned-el	42.3
	2	LC	7lang-ov-ft-avg	43.8
	3	LC	7lang-ov	44.1
		PROMT	multilingual-model-round2	44.3
		LC	7lang	44.8
	4	ETRANSLATION	1	46.2
		Baseline	Transformer+	46.9
	5	LC	1lang	47.3
	-	Baseline	Transformer	48.2
	6	CUNI-MT	transfer	51.6
	7	ETRANSLATION	2	53.2
	8	PROMT	multilingual-model-round1	55.2
	9	CUNI-MT	multiling	56.1
		CdT-ASL	only-round2-data	56.6
Unconst.	1	ETRANSLATION	2	43.9
		PROMT	Transformer	44.0
	2	ETRANSLATION	1	44.7
	3	CdT-ASL	only-cdt-data	50.0

Table 25. Results of the English–Modern Greek language pair, divided by categories. Systems are ranked according to BEER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	BEER [↑]
Constrained	1	PROMT	multilingual-model-round2-tuned-el	67.8
	2	LC	7lang-ov-ft-avg	67.2
	3	LC	7lang-ov	67.0
	4	LC	7lang	66.5
		PROMT	multilingual-model-round2	66.3
	5	ETRANSLATION	1	65.5
	6	LC	1lang	64.8
		Baseline	Transformer+	64.7
	-	Baseline	Transformer	63.7
	7	CUNI-MT	transfer	61.3
	8	ETRANSLATION	2	60.8
	9	CUNI-MT	multiling	59.5
		PROMT	multilingual-model-round1	59.5
		CdT-ASL	only-round2-data	59.0
Unconst.	1	PROMT	Transformer	67.2
		ETRANSLATION	2	66.9
	2	ETRANSLATION	1	66.3
	3	CdT-ASL	only-cdt-data	63.7

Table 26. Results of the English–Swedish language pair, 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.

	Rank	Team	Description	BLEU [↑]
Constrained	1	ETRANSLATION	4bigTens	22.7
		LC	5lang-ov-ft-avg	22.0
		PROMT	multilingual-model-round2-tuned-sv	21.8
		LC	5lang-ov-r2-data	21.8
	2	PROMT	multilingual-model-round2	20.4
	3	CdT-ASL	only-round2-data	20.3
	-	Baseline	Transformer+	19.5
	4	LC	7lang-ov-r1-data	18.3
	5	LC	5lang-r1-data	17.7
	6	PROMT	multilingual-model-round1	17.2
	7	LC	1lang-r1-data	16.7
		Baseline	Transformer	15.3
	8	CUNI-MT	multiling	14.7
		CUNI-MT	transfer	13.9
Unconst.	1	ETRANSLATION	4bigTens	23.3
	2	CdT-ASL	only-cdt-data	21.3
		PROMT	Transformer	21.0

Table 27. Results of the English–Swedish language pair, divided by categories. Systems are ranked according to TER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	TER [↓]
Constrained	1	PROMT	multilingual-model-round2-tuned-sv	69.3
	2	PROMT	multilingual-model-round2	70.7
	3	LC	5lang-ov-r2-data	71.5
		LC	5lang-ov-ft-avg	71.7
		Baseline	Transformer+	72.2
		ETRANSLATION	4bigTens	72.7
	4	LC	7lang-ov-r1-data	74.9
		CdT-ASL	only-round2-data	75.3
		PROMT	multilingual-model-round1	75.3
		LC	5lang-r1-data	75.6
	5	CUNI-MT	transfer	76.5
		Baseline	Transformer	77.5
	6	LC	1lang-r1-data	78.9
		CUNI-MT	multiling	79.3
Unconst.	1	ETRANSLATION	4bigTens	70.2
	2	PROMT	Transformer	71.3
	3	CdT-ASL	only-cdt-data	72.7

Table 28. Results of the English–Swedish language pair, divided by categories. Systems are ranked according to BEER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	BEER [↑]
Constrained	1	PROMT	multilingual-model-round2-tuned-sv	49.7
		LC	5lang-ov-r2-data	49.4
		LC	5lang-ov-ft-avg	49.2
	2	PROMT	multilingual-model-round2	48.9
	3	ETRANSLATION	4bigTens	48.2
		Baseline	Transformer+	48.1
	4	LC	7lang-ov-r1-data	47.4
	5	LC	5lang-r1-data	47.0
	6	PROMT	multilingual-model-round1	46.7
		CdT-ASL	only-round2-data	46.5
	7	LC	1lang-r1-data	45.4
	8	CUNI-MT	multiling	45.1
	-	Baseline	Transformer	44.4
	9	CUNI-MT	transfer	43.5
Unconst.	1	ETRANSLATION	4bigTens	50.0
	2	PROMT	Transformer	49.3
	3	CdT-ASL	only-cdt-data	48.7

Table 29. Results of the English–Arabic language pair, 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.

	Rank	Team	Description	BLEU [↑]
Constrained	1	LC	7lang-ov	25.1
	2	PROMT	multilingual-model-round2-tuned-ar	22.9
	3	LC	7lang	22.0
		PROMT	multilingual-model-round2	21.7
	4	CUNI-MT	transfer	19.1
		LC	1lang	19.1
		Baseline	Transformer	18.8
	5	CUNI-MT	multiling	17.0
	6	CdT-ASL	only-round2-data	15.9
Un.	1	PROMT	Transformer	31.4

Table 30. Results of the English–Arabic language pair, divided by categories. Systems are ranked according to TER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	TER [↓]
Constrained	1	PROMT	multilingual-model-round2-tuned-ar	62.9
	2	PROMT	multilingual-model-round2	63.6
	3	LC	7lang-ov	64.7
	4	LC	7lang	67.4
	5	CUNI-MT	transfer	68.7
		Baseline	Transformer	69.3
	6	LC	1lang	73.8
	7	CUNI-MT	multiling	75.2
	8	CdT-ASL	only-round2-data	77.9
Un.	1	PROMT	Transformer	54.2

Table 31. Results of the English–Arabic language pair, divided by categories. Systems are ranked according to BEER. Lines indicate clusters according to ART. Systems within a cluster are considered tied and, thus, are ranked equally.

	Rank	Team	Description	BEER [↑]
Constrained	1	LC	7lang-ov	57.6
	2	PROMT	multilingual-model-round2-tuned-ar	56.5
	3	PROMT	multilingual-model-round2	55.9
		LC	7lang	55.8
	4	LC	1lang	53.0
		CUNI-MT	transfer	52.9
	-	Baseline	Transformer	52.3
	5	CUNI-MT	multiling	51.3
	6	CdT-ASL	only-round2-data	48.7
Un.	1	PROMT	Transformer	62.3

4.8 Conclusions

This second round addressed 7 different language pairs and was divided into two categories: one in which participants were limited to using only the provided corpora (constrained) and other in which it the use of external tools and data was allowed (unconstrained).

References

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

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

[3] Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with LSTM. Neural computation **12**(10), 2451–2471 (2000)

[4] Jacquet, G., Verile, M.: Covid-19 news monitoring with medical information system (medisys). European Commission, Joint Research Centre (JRC) (2020)

[5] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., 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, Melbourne, Australia (July 2018), URL `http://www.aclweb.org/anthology/P18-4020`

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

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

[8] Kocmi, T., Bojar, O.: Trivial transfer learning for low-resource neural machine translation. pp. 244–252 (01 2018), https://doi.org/10.18653/v1/W18-6325

[9] Kudo, T., Richardson, J.: 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`

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

[11] Linge, J., Steinberger, R., Fuart, F., Bucci, S., Belyaeva, J., Gemo, M., Al-Khudhairy, D., Yangarber, R., van der Goot, E.: MediSys: medical information system. in advanced ICTs for disaster management and threat

detection: Collaborative and distributed frameworks. IGI Global pp. 131–142 (2010)
[12] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation (2020)
[13] Molchanov, A.: PROMT systems for WMT 2019 shared translation task. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 302–307, Association for Computational Linguistics, Florence, Italy (Aug 2019), https://doi.org/10.18653/v1/W19-5331, URL https://www.aclweb.org/anthology/W19-5331
[14] Oravecz, C., Bontcheva, K., Tihanyi, L., Kolovratnik, D., Bhaskar, B., Lardilleux, A., Klocek, S., Eisele, A.: etranslation's submissions to the wmt 2020 news translation task. In: Proceedings of the Fifth Conference on Machine Translation, pp. 254–261, Association for Computational Linguistics (2020)
[15] Papavassiliou, V., Prokopidis, P., Thurmair, G.: 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, pp. 43–51 (2013)
[16] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
[17] Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 392–395 (2015)
[18] Post, M.: A call for clarity in reporting bleu scores. In: Proceedings of the Third Conference on Machine Translation, pp. 186–191 (2018)
[19] Riezler, S., Maxwell, J.T.: 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, pp. 57–64 (2005)
[20] Roussis, D., Papavassiliou, V., Sofianopoulos, S., Prokopidis, P., Piperidis, S.: Constructing parallel corpora from covid-19 news using medisys metadata (2022), under review at LREC 2022.
[21] Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 1715–1725 (2016)
[22] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the Association for Machine Translation in the Americas, pp. 223–231 (2006)
[23] Stanojević, M., Sima'an, K.: Fitting sentence level translation evaluation with many dense features. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 202–206 (2014)

[24] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, vol. 27, pp. 3104–3112 (2014)
[25] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
[26] Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pp. 2214–2218, European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf
[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
[28] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., Liu, T.Y.: Incorporating bert into neural machine translation (2020)