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 participants systems are described showing improvements when using multilingual models sharing all the constrained data among the language pairs in constrained option. Similar systems but adding in-domain or using big corpora show best results in unconstrained option.

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.

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

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

(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. These systems are classified in the following 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 addresses the following language pairs:

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

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

3.2 Data Generation

Crawling Data Acquisition
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 [21] by harvesting the website of the European Medicines Agency[6], 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[7], 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://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 [12], 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

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

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.

Tests Selection
Given the set of documents obtained from the data crawling, for each language pair, we sorted all segments 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.

3.3 Corpora

The corpora was selected among the data generated in the previous section (see Section 3.2). Table 1 contains their statistics.

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

3.4 Evaluation

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

[8] `https://github.com/facebookresearch/LASER`.

BiLingual Evaluation Understudy (BLEU) [13]: geometric average of the modified n-gram precision, multiplied by a brevity factor.

Character n-gram F-score (ChrF) [14]: character n-gram precision and recall arithmetically averaged over all n-grams.

Among the two of them, BLEU was selected as the main metric and, thus, it was used to rank the participants.

We used `sacreBLEU` [15] in order to ensure consistent scores. Additionally, we applied Approximate Randomization Testing (ART) [16]—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[9].

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, 19] and another one based on the Transformer architecture [23]. All systems were built using `OpenNMT-py` [6].

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 [5], with a fixed learning rate of 0.0002 and a batch size of 60; label smoothing of 0.1 [20]; beam search with a beam size of 6; and joint byte pair encoding (BPE) [17] applied to all corpora, using 32,000 merge operations.

Similarly, systems for the Transformer baselines were trained using the standard parameters: 6 layers; Transformer [23], 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 [5], 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 [20]; beam search with a beam size of 6; and joint byte pair encoding (BPE) [17] applied to all corpora, using 32,000 merge operations.

3.6 Participants' approaches

PROMT The PROMT's approach consists in a multilingual model trained using MarianNMT[4] transformer architecture.

For constrained option, all data is concatenated using deduplication to one single multilingual corpus to build a 8k SentencePiece[8] 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 to the beginning of the English sentence of the English-Italian sentence pair).The author also removed all tokens that appear less than ten times in the combined deduplicated monolingual corpus from our vocabulary.

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

For unconstrained option, all available data mainly from the OPUS[22] and statmt[10] with the addition of private data harvested from the Internet was added to the training data. A special BPE implementation[11] 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 options. The PROMT system ranks the first but the English-German unconstrained option. The team plans to tune the baseline systems for the second round.

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

1. The standard NMT approach relies on one bidirectional model (sharing the encoder and decoder for both translation directions) which constantly switches between the training and the inference mode to produce batches of synthetic sentence pairs and learn from both authentic and synthetic training samples using online back-translation[9]. The models are trained on BPE units[17] with a vocabulary of 30k items.
2. The second consists of a transfer learning approach proposed by Kocmi and Bojar[7] (one of the participants) who fine-tune 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 replace the training data with the child corpus. They experiment 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 child's.
3. The multilingual 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.

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

The training was performed using the XLM[11] toolkit and the vocabulary size was set to 30k. The CUNI-MT system ranks the first for English-German and English-Swedish for constrained option.

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

For unconstrained systems, they use the UFAL Medical Corpus[12] for training data and then fine-tune models with constrained data.

All the data is tokenised using Khresmoi[13] tokeniser 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 single translation models for English to German and French language pairs.

They applied unigram SentencePiece for subword segmentation using shared vocabulary of source and target of 50K for single and 70K for multilingual models. Additionaly, 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 to German multilingual model achieves much higher score than the English to German single model and this improvement is not shown in the English to French model.

The LC system ranks the first for English-Swedish constrained option. For next rounds, they plan to use transfer learning from massive language models.

LIMSI LIMSI's team submitted systems to English to French constrained and unconstrained options. BPE using 32K vocabulary units was applied to the constrained system. For unconstrained systems, 4 systems were submitted: 1) one system build using additional in domain biomedical corpora, 2) a system first train on WMT14 general data and fine tuning on in-domain corpus, 3) same as 2 system but adding BERT[24] and 4) a system only trained with constrained data but BPE codes were computed using all the in-domain corpus.

The systems are trained using transformer architecture from *fairseq*[14] (Facebook's seq-2-seq library).

TARJAMA-AI Tarjama-AI team submitted systems for English to Spanish, German, Italian, French and Swedish constrained option. This system consists of a model trained with all the language pairs data adding a special token for the non target languages. Additionally, also oversamble the desired the target language.

[11] https://github.com/facebookresearch/XLM
[12] http://ufal.mff.cuni.cz/ufal_medical_corpus
[13] http://www.khresmoi.eu/assets/Deliverables/WP4/KhresmoiD412.pdf
[14] https://fairseq.readthedocs.io/en/latest/models.html

E-Translation E-Translation team submitted systems for English to German and English to French language pairs. They used transformer models from MarianNMT toolkit.

For English to German, they used transfer learning for constrained option and 12K size vocabulary created using SentencePiece. For unconstrained option, they submitted their WMT system and did fine tuning with contrained data.

They also participated in English-French constrained option with a small and big description system and English-French unconstrained option with *gen*, *phwt* and *eufl* description systems. E-Translation system ranks the first for English-German unconstrained option.

ACCENTURE This participants' report is missing but they annotated that they used multilingual BART[10] model in the description system.

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 outperform all others in lower ranking clusters, according to ART.

Constrained
Table 2 presents the results for constrained English–German. 12 different systems from 6 participants were submitted for this language pair. The best results 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 our baseline based on Transformer 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 tokens to all sentences except the ones from English–German). Next we have *CUNI-MTIR*'s Transformer approach. Finally, our RNN baseline and *TARJAMA-AI*'s NMT approach (they did not specify the architecture they used to train their system) placed last.

Table 3 presents the results for constrained English–French. 12 different systems from 8 participants were submitted for this language pair. *PROMT*'s multilingual approach yielded the best results. Second on the ranking are *ETRANSLATION*'s *small* and *big* approaches[15], *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 transfer learning approach, *CUNI-MTIR*'s Transformer

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

Table 2. Results of the constrained English–German language pair. 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 [↑]
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

approach, our RNN baseline, *TARJAMA-AI*'s NMT approach and *ACCENTURE*'s multilingual bart approach.

Table 3. Results of the constrained English–French language pair. 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 [↑]
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

Table 4 presents the results for constrained English–Spanish. 9 different systems from 6 participants were submitted for this language pair. *PROMT*'s multilingual approach yielded the best 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 positions we have *CUNI-MTIR*'s Transformer approach, *TARJAMA-AI*'s NMT approach and *ACCENTURE*'s multilingual bart approach (respectively).

Table 4. Results of the constrained English–Spanish language pair. 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 [↑]
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

Table 5 presents the results for constrained English–Italian. 5 different systems from 4 participants were submitted for this language pair. *PROMT*'s multilingual approach yielded the best 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.

Table 6 presents the results for constrained English–Modern Greek. 3 different systems from 2 participants were submitted for this language pair. 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 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 7 presents the results for constrained English–Swedish. 7 different systems from 5 participants were submitted for this language pair. The best results were yielded by *PROMT*'s multilingual approach, *LC*'s multilingual approach and one of *CUNI-MT*'s transfer learning approach. The other *CUNI-MT*'s trans-

Table 5. Results of the constrained English–Italian language pair. 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 [↑]
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

Table 6. Results of the constrained English–Modern Greek language pair. 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 [↑]
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

fer learning approach placed second on the ranking. Then 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.

Table 7. Results of the constrained English–Swedish language pair. 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 [↑]
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

Overall, multilingual and transfer learning approaches yielded the best results for all languages pairs. In fact, exept 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.

In general, differences from one position to the next one were few points (according to both metrics), with a case 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 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 (with a few exceptions in which there is a cluster above the RNN baselines of a similar translation quality).

Unconstrained

Table 8 presents the results for the unconstrained English–German language pair. 4 different systems from 3 participants were submitted for this language pair. The best results were achieved by *ETRANSLATION*'s WMT system fine-tuned with the in-domain data[16]. 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.

Table 8. Results of the unconstrained English–German language pair. 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 [↑]
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

Table 9 presents the results for the unconstrained English–French language pair. 8 different systems from 4 participants were submitted for this language pair. This is the language pair with most submissions for this category. The best 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

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

BPE using additional in-domain corpora. Then, we have *ETRANSLATION*'s *eufl* approach. Finally, we have *CUNI-MTIR*'s Transformer approach.

Table 9. Results of the unconstrained English–French language pair. 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 [↑]
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

Table 10 presents the results for the unconstrained English–Spanish language pair. Only 2 different systems from 2 participants were submitted for this language pair. The best results were achieved by *PROMT*'s multilingual system, followed by *CUNI-MTIR*'s Transformer approach.

Table 10. Results of the unconstrained English–Spanish language pair. 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 [↑]
1	PROMT	transformer	58.2	0.762
2	CUNI-MTIR	r1	32.1	0.582

Table 11 presents the results for the unconstrained English–Italian language pair. Only *PROMT*'s multilingual approach was submitted for this language pair.

Table 11. Results of the unconstrained English–Italian language pair. Systems are ranked according to BLEU.

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	transformer	38.0	0.642

Table 12 presents the results for the unconstrained English–Modern Greek language pair. Only *PROMT*'s multilingual approach was submitted for this language pair.

Table 12. Results of the unconstrained English–Modern Greek language pair. Systems are ranked according to BLEU.

Rank	Team	Description	BLEU [↑]	chrF [↑]
1	PROMT	transformer	42.4	0.652

Table 13 presents the results for the unconstrained English–Swedish language pair. Only 2 different systems from 2 participants were submitted for this language pair. The best results were achieved by *PROMT*'s multilingual system, followed by *CUNI-MTIR*'s Transformer approach.

Table 13. Results of the unconstrained English–Swedish language pair. 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 [↑]
1	PROMT	transformer	41.3	0.671
2	CUNI-MTIR	r1	24.0	0.514

Overall, this task had less participation than the constrained category. With the exemption of English–German—in which *ETRANSLATION* approaches based on WMT 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 external data.

3.8 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 Translation Error Rate (TER) [18]. This metric computes the number of errors between a translation hypothesis

and its reference (in this case, between the automatic reference and its post-edited version). Thus, the smallest the value the highest the quality. We obtained a TER value of 18.8, which is coherent with the translators opinion about the translations being generally good.

As a second step, we re-evaluated participant's translations (the corresponding subset only) using both the reference and its post-edited version. Table 14 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 3.7) are significant.

Table 14. 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. The PROMT's approach yielded the best results for all language pairs in both categories excepting for English-German unconstrained option and English-German sharing the first position with CUNI-MT.

In general, there are not big differences between ranked systems (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 next positions (excepting few cases where there are clusters above the RNN baselines of a similar translation quality).

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] 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`

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

[6] 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)

[7] 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

[8] 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`

[9] 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`

[10] 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)

[11] 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
[12] 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)
[13] 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)
[14] 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)
[15] Post, M.: A call for clarity in reporting bleu scores. In: Proceedings of the Third Conference on Machine Translation, pp. 186–191 (2018)
[16] 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)
[17] 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)
[18] 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)
[19] 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)
[20] 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)
[21] Tiedemann, J.: News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In: Recent Advances in Natural Language Processing, vol. V, pp. 237–248 (2009)
[22] 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
[23] 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)
[24] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., Liu, T.Y.: Incorporating bert into neural machine translation (2020)