Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez,
Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, Francisco Guzmán

Facebook AI

{guw,vishrav,angelafan,sahir}@fb.com

{namangoyal,somyaj,dkiela,tthrush,fguzman}@fb.com

Abstract

We present the results of the first task on Large-Scale Multilingual Machine Translation. The task consists on the many-to-many evaluation of a single model across a variety of source and target languages. This year, the task consisted on three different settings: (*i*) SMALL-TASK1 (Central/South-Eastern European Languages), (*ii*) the SMALL-TASK2 (South East Asian Languages), and (*iii*) FULL-TASK (all 101 x 100 language pairs). All the tasks used the FLORES-101 datasetas the evaluation benchmark. To ensure the longevity of the dataset, the test sets were not publicly released and the models were evaluated in a controlled environment on Dynabench.There were a total of 10 participating teams for the tasks, with a total of 151 intermediate model submissions and 13 final models. This year's result show a significant improvement over the known baselines with +17.8 BLEU for SMALL-TASK2, +10.6 for FULL-TASK and +9.4 for SMALL-TASK1.

1 Introduction

Despite recent advances in translation quality for a handful of languages and domains, MT systems still perform poorly on *low-resource languages*. Yet, most of the world's population speak low-resource languages and would benefit from improvements in translation quality on their native languages. As a result, the field has been shifting focus towards the evaluation of MT in low-resource situations (Thu et al., 2016; Guzmán et al., 2019; Barrault et al., 2020; ∀ et al., 2020; Ebrahimi et al., 2021; Kuwanto et al., 2021). However, these efforts have had poor coverage of low-resource languages which limits our understanding on generalization. More importantly, there has been little focus on the evaluation of true many-to-many multilingual models, which hampers the progress of the field despite all the recent excitement on this research direction (Fan et al., 2020).

The recent release of the FLORES-101 (Goyal et al., 2021) benchmark made possible to evaluate massively multilingual systems in a consistent way. The benchmark consists of 3001 sentences sampled from English Wikipedia and professionally translated in 101 languages. This poses a unique opportunity to understand translation across many languages with varied typology, resources, etc.

In this first multilingual large-scale shared task, we use the FLORES-101 benchmark to evaluate the progress on massively multilingual translation, where the evaluation is performed in a non-English-centric way. We propose 3 different tasks: two small tasks involving translation between 6 languages each (30 pairs), and a large task involving the translation across 101 languages (10K pairs). In the remainder of this paper, we describe the task setup, the participants, and the official results for the task. We also analyze the results to understand better the languages for which progress has been attained, and those where a gap in quality is still observed. Finally, we propose future directions for other tasks in the future.

2 Shared tasks

In this section, we briefly describe each of the tasks, the data, the baselines and metric used for evaluation.

2.1 Languages

The languages and statistics for the languages in the small tasks can be observed in Table 1, while the statistics for the complete set of languages in the full task can be obtained in Goyal et al. (2021).

SMALL-TASK1 - This task consisted of English and Central and South-Eastern European Languages: Croatian, Estonian, Hungarian, Macedonian, Serbian. These languages were chosen by their low availability of resources, geographical

ISO 639-3	Language	Family	Subgrouping	Script	Bitext w/ En	Mono Data
SMALL-TASK1						
hrv	**Croatian**	Indo-European	Balto-Slavic	Latin	42.2K	144M
est	**Estonian**	Uralic	Uralic	Latin	4.82M	46M
hun	**Hungarian**	Uralic	Uralic	Latin	16.3M	385M
mkd	**Macedonian**	Indo-European	Balto-Slavic	Cyrillic	1.13M	28.8M
srp	**Serbian**	Indo-European	Balto-Slavic	Cyrillic	7.01M	35.7M
SMALL-TASK2						
ind	**Indonesian**	Austronesian	Austronesian	Latin	39.1M	1.05B
jav	**Javanese**	Austronesian	Austronesian	Latin	1.49M	24.4M
msa	**Malay**	Austronesian	Austronesian	Latin	968K	77.5M
tam	**Tamil**	Dravidian	Dravidian	Tamil	992K	68.2M
tgl	**Filipino** (Tagalog)	Austronesian	Austronesian	Latin	70.6K	107M

Table 1: **Languages in each of the small tasks.** We include the ISO 639-3 code, the language family, and script. We also include the amount of resources available in OPUS as reported by Goyal et al. (2021)

proximity, language family diversity (Balto-Slavic, Uralic and Germanic), and different scripts.

SMALL-TASK2 This task consisted of English and South-Eastern Asian languages: Javanese, Indonesian, Malay, Filipino (Tagalog) and Tamil. These were chosen by their low-resource nature, geographical proximity and relatedness to a high-resource language (Indonesian).

FULL-TASK This task consisted of all 101 languages in the FLORES-101 benchmark, including English.

2.2 The evaluation data

The original sentences in FLORES-101 were sourced in English, from a broad group of topics that could be of general interest regardless of the native language of the reader. The sentences were sampled equally from *Wikinews*, *Wikijunior* and *WikiVoyage* by selecting an article randomly from each domain, and then selecting 3 to 5 contiguous sentences (not considering segments with very short or malformed sentences).

All source sentences were sent to a Language Service Provider (LSP) for translation into 101 languages. After that, the data was sent to different translators within the LSP for editing and quality assessment which then moved on to an automated quality control setup to ensure that the translation quality score was at least 90 on a scale of 0-100.

2.3 The baselines

Fan et al. (2021) worked on creating a Many-to-Many translation model, but it did not have the full coverage of languages in FLORES-101. Hence, we used the extended model trained in Goyal et al. (2021) which was supplemented with OPUS data and extended to 124 total languages. We trained two different sizes of models with 615M and 175M parameters.

2.4 Evaluation Metric

Automatically evaluating translation quality using BLEU is suboptimal as it relies on n-gram overlap which is heavily dependent on the particular tokenization used. The challenge of making BLEU comparable by using equivalent tokenization schemes has been partially addressed by `sacrebleu` (Post, 2018). Ideally, the automatic evaluation process should be robust, simple and can be applied to any language without the need to specify any particular tokenizer, as this will make it easier for researchers to compare against each other.

Towards this goal, we trained a SentencePiece (SPM) tokenizer (Kudo and Richardson, 2018) with 256K tokens using the CC100 monolingual data[1] (Conneau et al., 2020; Wenzek et al., 2020) from all the FLORES-101 languages. SPM is a system that learns subword units based on untokenized training data, providing a *universal* tokenizer that can operate on any language. One challenge is that the amount of monolingual data available for different languages is not the same — an effect that is extreme when considering low-resource languages. Languages with small quantities of data may not have the same level of coverage in subword units, or an insufficient quantity of sentences to represent a diverse enough set of content. To address

[1] http://data.statmt.org/cc-100/

this, we train our SPM model with temperature upsampling similar to Conneau et al. (2020), so that low-resource languages are represented. Finally, to compute BLEU, we apply SPM tokenization to the system output and the reference, and then calculate BLEU in the space of sentence-pieces. Due to the difference in tokenization, spBLEU scores are not strictly comparable across different target languages. However, to compare different models, here we use averages across the same set of target languages assuming that difference in tokenizations do not favor any specific model. In Goyal et al. (2021) this metric is described as spBLEU, but in this paper we use BLEU and spBLEU interchangeably.

3 Participants

In this section, we list each of the task participants and briefly describe each of their submissions. For reproducibility, we link to each of the model submitted, available in the Dynabench platform.

eBay (Liao et al., 2021) This submissions compares different kind of back-translation settings to improve the baseline model. They compare different generation algorithms: top-5 beam search; regular decoding without beam search; regular decoding with sampling from top-10 words. Contrary to Edunov et al. (2018), they find that top-10 decoding works best. They also consider how much English data should be used for the back translation (since it's more abundant than for the other languages). The models are trained from scratch using iterative back translation. **Models:** model 440 (SMALL-TASK1), model 441 (SMALL-TASK2), model 425 (FULL-TASK)

Huawei-TSC (Yu et al., 2021) The Huawei-TSC's team use a deep transformer encoder-decoder architecture (Sun et al., 2019), and focus their efforts on a combination of heuristics for data preprocessing, synthetic data generation, fine-tuning language-specific layers, and ensemble knowledge distillation. Compared to their baseline transformer on devtest, they get +2.8 BLEU from the synthetic data generation, +0.5 BLEU from layer fine tuning, and +0.8 BLEU from the ensemble knowledge distillation. **Models:** model 439 (SMALL-TASK2)

LMU (Lai et al., 2021) The LMU team's submission was based on a multilingual model, which were improved based on two techniques: (*i*) Tagged back-translation originating from bilingual models (+1.6 above back-translation coming from a multilingual)[2]; (*ii*) data selection w.r.t to the dev/devtest corpora following (Axelrod et al., 2011). **Models:** model 444 (SMALL-TASK1)

Maastricht University (Liu and Niehues, 2021) This submission trained a single multilingual Machine translation system by training on all 30 directions of track 2 languages. They mainly adapted the released pretrained M2M-100 model. They also did some data filtering to create a cleaner version of training corpus. Also they created synthetic pairs by taking parallel source to pivot language translation dataset and automatically translating pivot language sentences into target language, which gives 0.5 BLEU score improvement. They also tried similarity regularizer and language specific adapter weight which give 0.2 BLEU score gains overall. **Models:** model 445 (SMALL-TASK2)

Microsoft (Yang et al., 2021) The Microsoft team participated in all three tasks. The submission is based on the newly-released pretrained model DeltaLM (Ma et al., 2021a). The final submission to the shared task uses a mixture of direct and pivoted translation to improve the performance of individual directions, depending on whether the direct or pivoted models perform best. The mixture results in an improvement of +3.63 BLEU for the FULL-TASK, over their baseline architecture (24/12), but smaller improvements for the SMALL-TASK2. In addition, the models use progressive learning, which starts with a smaller architecture, noisier training data, and later changes to improve performance. The model also uses a combination of parallel, back-translated and noisy-parallel data (obtained for langs. X and Y from back-translating into X and Y) **Models:** model 483 (FULL-TASK) model 448 (SMALL-TASK1) model 457 (SMALL-TASK2)

MMTAfrica (Emezue and Dossou, 2021) This submission creates a non-English-centric multilingual translation system focusing on six African languages (Igbo, Kinyarwanda, Fon, Swahili, Xhosa, Yoruba) and English and French. The system starts from mT5 (Xue et al., 2021) and finetunes it on parallel data with additional monolingual data used

[2] Authors hypothesized that the difference in performance could be due to the implicit *self-training* coming from a multilingual model, as opposed to the diversity introduced by a bilingual model.

for online backtranslation (Sennrich et al., 2016). To cover Fon and Kinyarwanda, which are not included in FLORES-101, a small new test set was created. Compared to the small baseline models provided in Goyal et al. (2021), significant improvements were obtained.

Samsung RPH - Konvergen AI (Sutawika and Cruz, 2021) The submission of the Samsung Research Philippines/Kovergen AI's team focuses on the languages in SMALL-TASK2, in particular on data preprocessing. For large-scale multilingual models, the importance of preprocessing has risen as researchers focus on using web crawls or noisily aligned data to train translation models. In this submission, various different preprocessing techniques are applied while holding the model and architecture fixed. The authors have gains of more than 1 BLEU point from improving preprocessing. **Models:** model 443 (SMALL-TASK2)

TenTrans (Xie et al., 2021) The submission explores several techniques to improve performance. It focuses on two systems: TenTrans and FLORES101, although the second one is favored in later experimentation. The authors achieve large improvements in performance by using a the pretrained M2M124 FLORES101 model. Main benefit comes from in-domain knowledge adaptation and fine-tuning. The authors use a domain classifier based on BERT. Then they use gradual fine-tuning to gradually removing the least-likely in-domain sentence pairs at the later stages of training. They also explore other techniques, including model averaging that help to improve the performance of their system. **Models:** model 460 (SMALL-TASK2)

TelU-KU (Budiwati et al., 2021) The team from TelU-KU participated in SMALL-TASK2. Their approach explores an interesting alternative of improving NMT performance via hyper-parameter optimization (most promising for low resource languages). Although simple, this approach effectively provides improvements by +1.08 BLEU on top of the small baseline and opens up a promising direction for hyper-parameter optimization. **Models:** model 465 (SMALL-TASK2)

UMD (Bandyopadhyay et al., 2021) This system build upon the baseline M2M-124 model (Fan et al., 2020). It includes two improvements: (*i*) fine-tuning over MultiCCAligned; (*ii*) it uses ReLUs, which improve +0.8 BLEU over GELUs. In addition, the final system is the result of an extensive hyper-parameter optimization. Interestingly, the authors find that using the bible for finetuning improves performance over the baseline model despite its small size (only about 0.5 BLEU behind MultiCCAligned). **Models:** model 304 (SMALL-TASK2)

4 Evaluation Environment

All models were evaluated within the Dynaboard evaluation-as-a-service framework (Ma et al., 2021b) that is a part of the Dynabench platform (Kiela et al., 2021). This was done to ensure that the FLORES test set remains hidden while we evaluate many-to-many translation. Moreover, the testing conditions were constrained to a p2.xlarge AWS instance, which has one NVIDIA K80 GPU.

All model submissions had to be wrapped in a torchserve[3] handler and were required to follow a fixed input/output specification using Dynalab[4]. Submitting a system to the task required writing some wrapper code, and often testing different configurations (e.g. batch size), to ensure that the model was able to run under the constraints.

Given the additional work needed to run the evaluation, participants were encouraged to test the platform and to submit models early on. To avoid fine-tuning on the devtest set, we established a submission cap of one model per day.

In total, we had 81 distinct model submissions for the small task2 (South-East Asian Languages), 57 distinct submissions to the small task1 (Central / South-East European Languages), and 13 model submissions to the full task. During the evaluation period, participants were requested to mark a model as their final submission. In the end, we had 10 final submissions to the small task2, 4 to the small task 1 and 3 to the full task.

In Figure 1 we observe the total number of submissions per day. We can see that the total number of submissions per day remained low (less than 5) until August, where the number of submissions reached 16 per day.

5 Results

Present the analysis of the results for each of the tasks. Furthermore, we analyze the progress made for each task, that is, how much improvement has

[3]https://pytorch.org/serve

[4]https://github.com/facebookresearch/dynalab

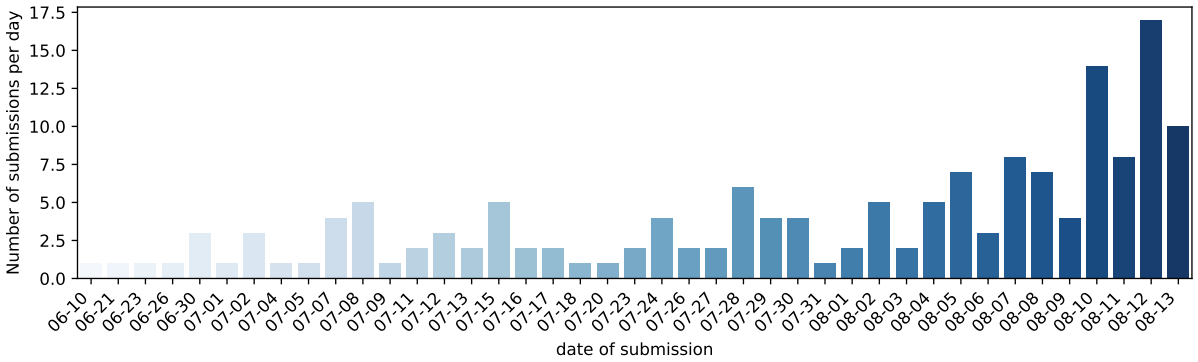

Figure 1: Submissions to the shared task through Dynabench per day. As expected, we see a rise in the number of submissions towards the end of the evaluation period.

there been between the baselines and the best models. Lastly, we analyze the difference between the models for the full task and each of the smaller tasks.

5.1 Main Results

In Table 2 we observe the final results for each of the shared tasks. From the results we observe that the DeltaLM model from the Microsoft team performs best by a large margin on the SMALL-TASK1 (+2.6 BLEU) and FULL-TASK (+9.1 BLEU), but the margin is smaller for the SMALL-TASK2 (0.6 BLEU). Below, analyze each task's results independently.

	BLEU
SMALL-TASK1 ***(CSE European langs)***	
Microsoft	37.59
eBay	34.96
LMU	31.86
baseline M2M-615	28.23
baseline M2M-175	21.33
SMALL-TASK2 ***(SE Asian langs)***	
Microsoft	33.89
eBay	33.34
TenTrans	28.89
Maastricht University	28.64
Huawei-TSC	28.40
Samsung RPH/ Konvergen AI	22.97
baseline M2M-615	16.11
UMD	15.72
TelU-KU	13.19
baseline M2M-175	12.30
FULL-TASK ***(all langs)***	
Microsoft	16.63
eBay	7.55
baseline M2M-175	6.05

Table 2: Official results for the three shared tasks in the large-scale multilingual machine translation task

SMALL-TASK1 In the Central/South-East European languages we observed that the model pre-trained with DeltaLM performed best, followed by eBay's model by a margin of 2.6 BLEU. In this task we observe that the progress between the M2M-615 baseline and the next best system of 3.6 BLEU.

SMALL-TASK2 In the South-East Asian languages task, there were many more submissions than in the other tasks. We see a smaller gap between the first and second models. These two models are very different, one using a large pre-trained language model, while the other one trains from scratch and uses iterative back translation. There is also a second cluster formed by the submission of the next three models, with a gap less than 0.5 BLEU among them. In this cluster, two models are based on the pre-trained M2M model while the third one is trained from scratch. Six out of eight participants perform better than the M2M-615 baseline, while all participants perform better than the M2M-175 baseline. The gap between the best system and the M2M-615 baseline is of 17.8 BLEU.

FULL-TASK In the full task we had fewer submissions, possibly due to the difficulty and resources to train an evaluate such models. Here the gap between the best and second-best models is significant, around 9 BLEU. However, note that the gap between the best systems and the baseline is much smaller ($\sim$10.6 BLEU), denoting how much harder is translating more languages with similarly sized models.

5.2 Analysis of the progress on quality

One interesting aspect that we can analyze is how much progress has there been since the release of M2M-100 (Fan et al., 2020), and its subsequent adaptation for FLORES101, M2M-124. Here, we break down the improvements by language pairs to understand better the changes in performance.

Note that looking at spm-BLEU numbers across target languages can be deceiving. This is due to the different spm vocabularies that are used for each target language. However, for the sake of simplicity in the following analyses we assume that: (*i*) relative improvements (deltas) are comparable across language pairs, (*ii*) averages of relative improvements from two different source languages (say English and Hausa) into the remaining 101 languages are roughly comparable, even though the average for Hausa on the source doesn't contain on the target Hausa and contains English, and the average for English on the source doesn't contain English on the target but contains Hausa.

5.2.1 Progress on SMALL-TASK1

SMALL-TASK1 is constrained and encompasses Central and South-East European Languages. In Table 3 we see that the top performing pairs (most progress) are into and out of English, while the worst performing ones include Croatian and Macedonian. The gap between the best and the worst performing pairs is of 13 BLEU, yet on average, translation across language pairs improved 11.3 BLEU.

Source	Target	ΔBLEU
Best 5		
English	Serbian	19.08
Serbian	English	15.58
Macedonian	English	14.81
Estonian	English	14.17
Hungarian	English	13.37
Worst 5		
Hungarian	Croatian	9.05
Macedonian	Croatian	8.09
Croatian	Macedonian	6.96
Serbian	Macedonian	6.49
Serbian	Croatian	6.13
Average:		11.32

Table 3: Progress in quality for the best and worst language pairs in SMALL-TASK1

In Table 4 we present the average progress for languages in the source or target, and we observe the following: there was more progress in translating from English than any other language. However, the gap between the best and worst is less than 5 BLEU. When looking at the performance when translating into each of the task languages, we see a very similar tendency: English tops the list, Croatian is at the bottom, and the gap between best performing and worst performing languages is less than 5 BLEU.

Source	ΔBLEU	Target	ΔBLEU
English	14.20	English	13.97
Macedonian	11.65	Serbian	13.58
Estonian	11.43	Hungarian	10.96
Hungarian	11.22	Estonian	10.91
Serbian	9.84	Macedonian	9.47
Croatian	9.58	Croatian	9.02

Table 4: Average progress for each of the languages in SMALL-TASK1

5.3 Progress on SMALL-TASK2

For SMALL-TASK2, there was a significant progress on languages like Tamil (tam) and Tagalog (tgl). In Table 5 we see a progress of 30+ BLEU for translation between Tamil <> English. This is encouraging, as the baseline model had issues translating from/into Tamil. It is also encouraging to see that even for the translation between Malay <> Indonesian (which was strong to begin with), we see more than 10+ BLEU improvement. On average, we see an improvement of 21.8 across all directions. It's important to note the fact that all submissions for this task were constrained, so these improvements come from better modeling and training techniques.

Another aspect to note comes from Table 6, where we see that the language with most progress is Tamil, followed by English and Tagalog. On the other hand, in this case we see more disparity on the progress between the languages with most and least progress. For instance, it is harder to translate into Javanese, which only improves 14.7 BLEU on average.

5.3.1 Progres on FULL-TASK

In Table 7 we present the deltas between the best scores in the competition for each language pair, and the baseline. We observe that there are significant improvements for certain languages, particularly: Welsh (cym), Irish (gle), Maltese (mlt) and their pairings with English. These are languages for which the original M2M model was doing poorly,

Source	Target	ΔBLEU
Best 5		
English	Tamil	32.63
English	Tagalog	31.04
Tagalog	English	30.16
Tamil	English	30.00
Indonesian	Tamil	28.45
Worst 5		
Tagalog	Javanese	14.67
Malay	Javanese	12.40
Indonesian	Malay	11.59
Indonesian	Javanese	11.05
Malay	Indonesian	10.45
Average:		21.75

Table 5: Progress in quality for the best and worst language pairs in SMALL-TASK2

Source	ΔBLEU	Target	ΔBLEU
Tamil	24.35	Tamil	27.29
English	24.30	Tagalog	25.29
Tagalog	23.19	English	24.68
Javanese	20.68	Malay	19.72
Indonesian	19.13	Indonesian	18.81
Malay	18.88	Malay	14.74

Table 6: Average progress for each of the languages in SMALL-TASK2

yet the DeltaLM model is doing much better [5]. In fact, as seen in Fig. 2, these language pairs are an exception, and most language pairs fall around the 11 BLEU improvement range. The average improvement across language pairs is 10.6 BLEU. However, there are several language pairs for which there was no progress at all. In Fig. 2, close to 10% (∼1K pairs) have less than 5 BLEU improvement.

[5]Since this is an unconstrained submission, it is hard to know what data went into the models. However, we hypothesize that the improvement is likely due to the amount of training data available for DeltaLM. As pointed out in Yang et al. (2021) their model contains about 300K sentences for Maltese (mt), 1.5M sentences for Irish (ga), and 3M sentences for Welsh (cy)

Source	Target	ΔBLEU
Best 5		
English	Welsh	46.41
Irish	English	43.55
English	Irish	43.10
Maltese	Welsh	42.88
Irish	Maltese	41.83
Average:		10.60

Table 7: Progress in quality for the best and worst language pairs in FULL-TASK. Note that we exclude the worst performing pairs, which made no progress at all.

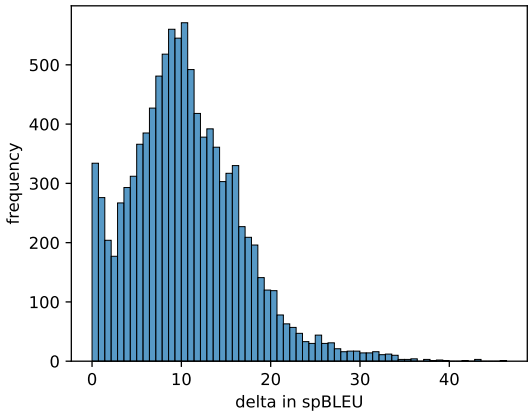

Figure 2: Distribution of improvements in BLEU for different language pairs in the full task

To facilitate the analysis of the progress across languages, in Fig. 3 we present the improvements by language groupings. We see big improvements coming from Other Indo-European (influenced by Irish, Welsh), Dravidian (influenced by Tamil, Telugu), Austronesian (influenced by Tagalog). However we note that there is very little progress for African Languages as represented by the Bantu and Nilotic subgroups. Another interesting finding is that progress trends to be lower when translating into harder languages.

In summary, there is large progress for a few languages, but sadly, there is little progress made for very low-resource languages, particularly those unrelated to other major languages.

5.4 Moderately Multilingual vs. Massively Multilingual

A natural question that arises is: what is the gap that remains between what we're calling moderately multilingual models (MoM), i.e models handle just a few languages and a couple dozen pairs; vs. the massively multilingual models (M2M) that handle hundreds of languages and tens of thousand pairs?

To analyze this aspect, we compare the best models for the full task, and each of the small tasks.

5.4.1 SMALL-TASK1 vs. FULL-TASK

In Figure 4 we present the scores of the best system for task1 (MoM) vs. the best system for the full task (M2M). Here we observe that there is a consistent gap of about 4.7 BLEU between the MoM and the M2M models when averaging across source languages. We can observe on the distribution of deltas of performance that drops in performance are similarly distributed across languages. This

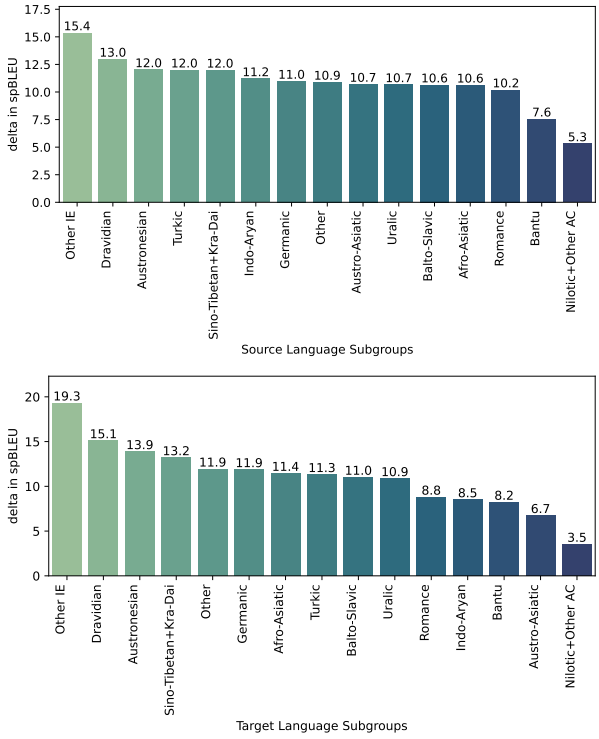

Figure 3: Average BLEU improvements per languages in the source and target language families

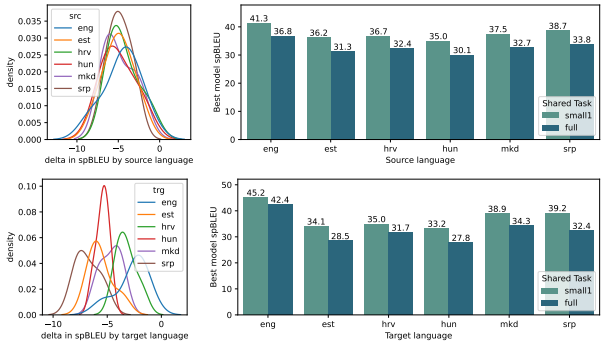

Figure 4: Comparison of average performances of the best systems in the FULL-TASK and SMALL-TASK1 by source and target languages

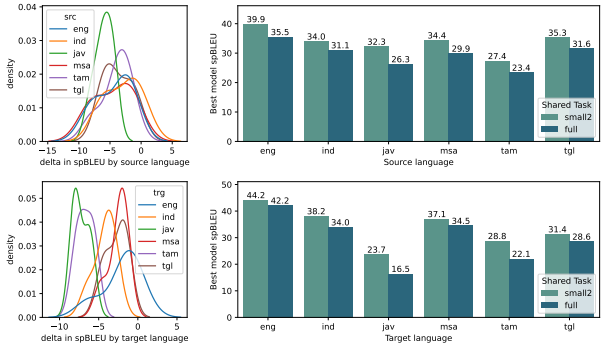

Figure 5: Comparison of average performances of the best systems in the FULL-TASK and SMALL-TASK2 by source and target languages

suggests that the *curse of multilinguality* (Conneau et al., 2020), i.e. the loss in performance by adding more languages into to a model with fixed capacity, affects equally the encoding of different languages to a rate of about 0.05 BLEU per language added to the model. This is encouraging, as it suggests that encoding is robust to the addition of new languages.

On the other hand, when we look at the target side the picture is quite different. Particularly, we observe more variation in performance, ranging from -2.7 BLEU for English to -6.8 BLEU for Serbian. We hypothesize that these differences could be due to a combination of factors: (*i*) amount of supervision (which would explain why English performance doesn't drop as much), (*ii*) additional supervision from similar languages, (*iii*) morphological richness (which would explain why Hungarian and Estonian are more affected), and (*iv*) script usage (which would explain why Serbian is more affected than Croatian). However, proving these hypotheses is beyond the scope of this paper.

5.4.2 SMALL-TASK2 vs. FULL-TASK

In Figure 5 we present the scores of the best system for task2 (MoM) vs. the best system for the full task (M2M). Here we see again that the model with more parameters per language is still ahead by about 4.26 BLEU. We also observe more variability in the distribution of drops in performance, notably, Javanese, the lowest resource language, being the most different to the others.

On the target side, we observe that English is ahead of the curve, showing the least regression. On the other hand Javanese and Tamil further reinforce our observations that additional supervision and morphology play an important role in decoding performance.

5.5 African Languages

While not officially a track on this year's competition, Emezue and Dossou (2021) focused on the task of multilingual machine translation for African languages that are in FLORES-101. They introduced MMTAfrica, the first many-to-many multilingual translation system for six African languages: Fon (fon), Igbo (ibo), Kinyarwanda (kin), Swahili/Kiswahili (swa), Xhosa (xho), and Yoruba (yor) and two non-African languages: English (eng) and French (fra). For multilingual translation concerning African languages, a novel backtranslation

and reconstruction objective, BT&REC, was introduced which is inspired by the random online back translation and T5 modelling framework respectively, to effectively leverage monolingual data. Additionally, MMTAfrica improves over the FLORES 101 benchmarks (BLEU gains ranging from +0.58 in Swahili to French to +19.46 in French to Xhosa).

6 Conclusion and Future Work

In this paper we presented the first iteration of the large-scale multilingual translation task. This task attracted several teams from across the globe and many models submissions. We kept the test set blind and used a platform to evaluate model submissions under a controlled environment. In this task, we observed significant progress in translation quality across tasks, but particularly in the *small* tasks. We observed that pre-trained language models and large amounts of back-translation (either at one go, or in iterative fashion) were important tools used by many participants.

We observed that models that have to translate fewer languages trend to do better on average, and that lower resources and morphology complicate translation, particularly for decoding. We also observed that languages in certain groups, like the African languages in the Bantu and Nilotic families, experience little to no improvement.

In the future, we want to organize shared tasks with languages for which little or no progress was achieved this time around. Additionally, we want to open up the FLORES evaluation setup to other organizers interested groups of languages within the FLORES-101 set.

Acknowledgements

We would like to thank Geeta Chouhan and Hamid Shojanazeri for their support setting up the GPU inference and batch decoding with torchserve. We would like to thank Carlos Escapa for his support in getting compute credits for this competition, and Microsoft Azure, Google Cloud and Amazon AWS for donating credits for participants.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Saptarashmi Bandyopadhyay, Tasnim Kabir, Zizhen Lian, and Marine Carpuat. 2021. The University of Maryland, College Park Submission to Large-Scale Multilingual Shared Task at WMT 2021. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Sari Dewi Budiwati, Tirana Fatyanosa, Mahendra Data, Dedy Rahman Wijaya, Patrick Adolf Telnoni, Arie Ardiyanti Suryani, Agus Pratondo, and Masayoshi Aritsugi. 2021. To Optimize, or Not to Optimize, That Is the Question: TelU-KU Models for WMT21 Large-Scale Multilingual Machine Translation. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, et al. 2021. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Conference of the Association for Computational Linguistics (ACL)*.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. 2021. MMTAfrica: Multilingual Machine Translation for African Languages. In *Proceedings of the*

Sixth Conference on Machine Translation, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in nlp.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Garry Kuwanto, Afra Feyza Akyürek, Isidora Chara Tourni, Siyang Li, and Derry Wijaya. 2021. Low-resource machine translation for low-resource languages: Leveraging comparable data, code-switching and compute resources. *arXiv preprint arXiv:2103.13272*.

Wen Lai, Jindřich Libovický, and Alexander Fraser. 2021. The LMU munich system for the wmt 2021 large-scale multilingual machine translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Baohao Liao, Shahram Khadivi, and Sanjika Hewavitharana. 2021. Back-translation for Large-Scale Multilingual Machine Translation. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2021. Maastricht University's Large-Scale Multilingual Machine Translation System for WMT 2021. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021a. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *CoRR*, abs/2106.13736.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021b. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. *Conference of the Association for Computational Linguistics (ACL)*.

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Baidu neural machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 374–381, Florence, Italy. Association for Computational Linguistics.

Lintang Sutawika and Jan Christian Blaise Cruz. 2021. Data Processing Matters: SRPH-Konvergen AI's Machine Translation System for WMT'21. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the Asian language treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1574–1578, Portorož, Slovenia. European Language Resources Association (ELRA).

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Wanying Xie, Bojie Hu, Han Yang, Dong Yu, and Qi Ju. 2021. TenTrans Large-Scale Multilingual Machine Translation System for WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei. 2021. Multilingual Machine Translation Systems from Microsoft for WMT21 Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Zhengzhe Yu, Daimeng Wei, Zongyao Li, Hengchao Shang, Xiaoyu Chen, Zhanglin Wu, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, Hao Yang, and Ying Qin. 2021. HW-TSC's Participation in the WMT 2021 Large-Scale Multilingual Translation Task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.