The OPUS-MT Dashboard – A Toolkit for a Systematic Evaluation of Open Machine Translation Models

Jörg Tiedemann and **Ona de Gibert**
Department of Digital Humanities
University of Helsinki, Helsinki / Finland
{jorg.tiedemann, ona.degibert}@helsinki.fi

Abstract

The OPUS-MT dashboard is a web-based platform that provides a comprehensive overview of open translation models. We focus on a systematic collection of benchmark results with verifiable translation performance and large coverage in terms of languages and domains. We provide results for in-house OPUS-MT and Tatoeba models as well as external models from the Huggingface repository and user-contributed translations. The functionalities of the evaluation tool include summaries of benchmarks for over 2,300 models covering 4,560 language directions and 294 languages, as well as the inspection of predicted translations against their human reference. We focus on centralization, reproducibility and coverage of MT evaluation combined with scalability. The dashboard can be accessed live at https://opus.nlpl.eu/dashboard/.

1 Introduction

The main motivation behind the OPUS-MT dashboard is to provide a comprehensive overview of open translation models. We focus on a systematic collection of benchmark results with verifiable translation performance and large coverage in terms of languages and domains. The landscape of Machine Translation (MT) is increasingly blurry and incomprehensible due to the growing volume of shared tasks and models published within the community. Even with established events such as the Conference on Machine Translation (WMT), a complete picture of translation performance is hard to obtain. In addition, large multilingual language and translation models push the language coverage making it difficult to keep an eye on the state of the art for particular language pairs.

One additional problem is that most results reported in scientific and non-scientific channels come from selected benchmarks and model performance and are not explicitly verified by a careful replication study. In various cases, new models come with their own benchmarks and do not consider a wider evaluation across domains. Training data is complicated to control and the danger of over-fitting to specific scenarios is apparent. Figure 1 illustrates the substantial differences one can observe across benchmarks when comparing two competing models.

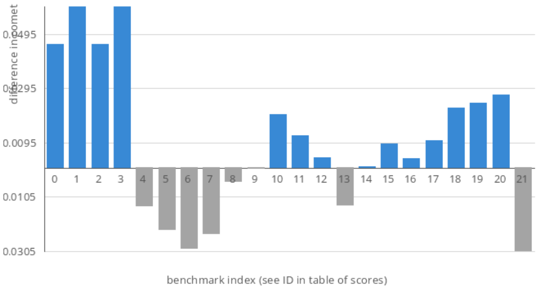


Figure 1: Difference between two models in terms of COMET scores across 21 benchmarks for English-to-French translation.

Our dashboard is an attempt to carefully provide a summary of results using an extendable collection of benchmarks with the largest language coverage possible accommodated with procedures to translate and evaluate in a standardized and consistent setup. The focus is clearly set on publicly available translation models as we want to emphasize translation results that we can replicate and verify from our own experience. The system is designed with the following requirements in mind:

- comprehensive and scalable collection
- lean and responsive interface
- open and transparent implementation

The implementation and all data files are available on GitHub in public repositories and details about the components and implementations are given below. We start by a brief description of the background before discussing the collection of benchmarks and translation evaluations. The main features of the web interface are listed thereafter

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 315–327
July 10-12, 2023 ©2023 Association for Computational Linguistics

and we finish up with links to related work and an outlook into future developments.

2 Background and Motivation

The main motivation of the dashboard is related to our own initiative on building open translation models under the umbrella of OPUS-MT. The development of MT accelerated in recent years making it difficult to obtain a clear view on performance for individual language pairs. In contrast to related work, the dashboard is not intended to serve as a new MT evaluation service but rather as a central point of comparison between OPUS-MT and other publicly available models. User-provided translations are also supported as another point of reference but the focus is set on verifiable translation results produced by the system itself.

OPUS-MT is based on OPUS (Tiedemann, 2012), the major hub of public parallel data, the main ingredient for training modern translation models. It refers to an initiative to systematically train translation models on open data sets using the efficient NMT framework provided by Marian-NMT (Junczys-Dowmunt et al., 2018). The project includes both bilingual and multilingual transformer models of different sizes with streamlined procedures to scale up language coverage and availability (Tiedemann and Thottingal, 2020). OPUS-MT models are released with permissive licenses and can easily be integrated in existing workflows (Tiedemann et al., 2022; Nieminen, 2021; Tiedemann et al., 2023). Currently, there are over 2,300 public models – release information is part of the dashboard.[1] The models cover over 290 languages and provide translations for 2,549 language pairs and 4,560 language directions. The sheer volume of OPUS-MT releases call for a systematic evaluation to monitor progress and support model selection in practical applications.

3 Collecting MT Benchmarks

MT benchmarks are developed for various tasks and domains and their distribution differs depending on the preferences of the original provider. In order to make it easier to systematically compare MT across available benchmarks, we collect known testsets in a unified and consistent format in a public repository[2] (OPUS-MT-testsets).

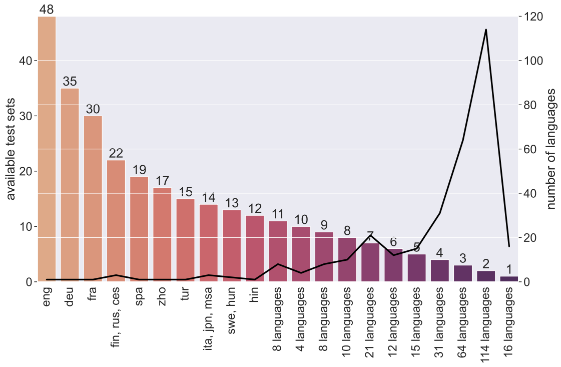


Figure 2: Distribution of languages covered in public benchmarks bucketed by the number of available test sets per language. The solid line shows the number of languages in each bucket.

The current collection covers benchmarks from translation tasks at WMT (Kocmi et al., 2022), TICO19 (Anastasopoulos et al., 2020), the Tatoeba translation challenge (Tiedemann, 2020), Flores v1.0 (Guzmán et al., 2019), Flores-101 (Goyal et al., 2022) and Flores-200 (NLLB Team et al., 2022a) and multi30k (Elliott et al., 2016). Each benchmark may be comprised of several testsets (different years or dev and test sets). Altogether we cover over 44,000 language directions with an average length of 1,945 sentences per testset. One important thing to note is that the multilingual benchmarks are typically English-centric in a sense that the original text has been translated from English to other languages. Figure 2 illustrates the skewed distribution and we encourage suggestions[3] and developments of additional sources to change that picture.

We sort benchmarks by language pair using ISO-639-3 language codes and use a simple plain text format with UTF-8 encoded data. Translated segments (typically sentences) appear on the same line in separate files for the input text in the source language and reference translations in the target language. The file name corresponds to the benchmark name and the file extension specifies the language by its ISO code. Additional files may provide additional metadata such as domain labels or source information. We also add extensions about the writing script if necessary. Here are a few examples from the collection:

```
eng-deu/newstest2020.deu
eng-deu/newstest2020.eng
```

[1] https://opus.nlpl.eu/dashboard/releases.php

[2] https://github.com/Helsinki-NLP/OPUS-MT-testsets/

[3] Suggestions can be proposed by adding issues to our GitHub repository.

```
srp_Cyrl-fin/flores200-devtest.fin
srp_Cyrl-fin/flores200-devtest.srp_Cyrl
```

Currently, we do not support multi-reference benchmarks. Instead, additional reference translations are stored as separate benchmarks. Document boundaries may be specified by empty lines. Other formatting information is not supported.

We will extend the repository with additional test sets including NTREX (Federmann et al., 2022), Samanantar (Ramesh et al., 2021), IWSLT (Antonios et al., 2022) and data from the Masakhane project (Orife et al., 2020).

4 Systematic Evaluation of Public Models

We store the results of our systematic evaluation in three different public git repositories, depending on the type of model: (i) Opus-MT models[4], (ii) external models[5], and (iii) user-contributed translations[6].

We emphasize a lean design avoiding the hassles of setting up and maintaining databases and additional services. Each leaderboard repository follows the same structure and is divided into two main parts: (i) leaderboards for each benchmark and language pair and (ii) the scores for each individual model. File structures are organized accordingly and the setup makes it possible to easily scale the collection to a large number of models, benchmarks and language pairs. The inclusion of new evaluation benchmarks is also straightforward as we have separate files for each of them. The main file structure looks like this:

```
scores/<src>-<trg>/<test>/<metric>-scores.txt
models/<org>/<model>/<test>.<metric>-scores.txt
```

Source and target language IDs (<src>, <trg>) and the name of the benchmark (<test>) correspond to the naming conventions used in OPUS-MT-testsets. Supported metrics are currently COMET (Rei et al., 2020),[7] BLEU (Papineni et al., 2002) and chrF (Popović, 2015) with the two variants chrF++ (Popović, 2017) and sentence-piece-based subword BLEU (Goyal et al., 2022). We use sacrebleu (Post, 2018) and unbabel-comet[8] to compute the results. We add the readily available Chinese, Japanese and Korean tokenization features in sacrebleu and the Flores-200 sentence piece model for subword splitting. Models are sorted by provider (<org>); and the model name (<model>) may be split into sub-directories specifying additional properties of the model. For example, OPUS-MT models are grouped by languages they support, which may be a single language pair or pairs of language groups.

Besides benchmark-specific score tables, we also compile aggregated score tables for each language pair. Those tables list an average score over several available benchmarks (avg-<metric>-scores.txt) and the best-performing model for each available benchmark (top-<metric>-scores.txt). Automatic makefile recipes are used to update those tables if needed. We keep separate tables for the three categories (OPUS-MT, external models, user-contributed translations) in each of the respective repositories.

Furthermore, the repository includes the procedures for translating and evaluating models with respect to the benchmarks collected as described in the previous section. The translations are systematically run with the same hyperparameters. These can be consulted in Appendix A. The implementation streamlines the creation of batch jobs and enables a scalable approach that allows a straightforward update of leaderboards with new models and benchmarks. Additional evaluation metrics can also be integrated by implementing appropriate recipes.

The general workflow for evaluating models and updating score tables is divided into three steps: (i) translating and evaluating all benchmarks for all language pairs supported by a model, (ii) registering new scores to be added to existing leaderboards, and (iii) updating all affected leaderboards and sorting by score. GNU makefile recipes are used to properly handle dependencies. Using revision control as the backend storage makes it possible to recover from errors and mistakes.

We distinguish between internal and external models but the basic workflow is the same. We anticipate that the publication of this paper will attract some interest allowing us to harvest additional user-contributed translations. We report statistics in Table 1. The relatively small amount of external models and the high number of translations is due to the fact that some of the external models

[4] https://github.com/Helsinki-NLP/OPUS-MT-leaderboard/

[5] https://github.com/Helsinki-NLP/External-MT-leaderboard/

[6] https://github.com/Helsinki-NLP/Contributed-MT-leaderboard/

[7] COMET model: Unbabel/wmt20-comet-da

[8] https://pypi.org/project/unbabel-comet

	OPUS-MT	Tatoeba	External
Providers	-	-	34
Models	693	1,633	78
Bilingual	634	779	73
Multilingual	59	854	5
Lang pairs/model	1.93	11.00	36.42
Translations	5,483	80,295	11,723
Testsets	49	66	49
Language directions	650	4,306	786
Language pairs	372	2,388	504
Language coverage	99	276	219

Table 1: Statistics on OPUS-MT, Tatoeba and external models.

are highly multilingual, in comparison, OPUS-MT models are mostly bilingual. Some additional details are given below.

4.1 OPUS-MT Models

OPUS-MT models are released as self-contained Marian-NMT models. They come in two flavors: Models trained on different selections from OPUS (Tiedemann and Thottingal, 2020)[9] and models trained on OPUS data released as part of the Tatoeba translation challenge (Tiedemann, 2020).[10] Release information is available from GitHub and feeds directly into the evaluation workflow. An update to the collection triggers new evaluation jobs. Missing benchmark scores can also be added using the dependency chains implemented in the leaderboard workflow. The actual jobs still need to be started manually as we have to control compute time allocations on the infrastructure that we employ. Translations and evaluations are typically done on some high-performance cluster (scheduled using SLURM (Yoo et al., 2003)) and we integrated the OPUS-MT leaderboard in the environment provided by CSC, the center for scientific computing in Finland.[11] However, all jobs should execute in any environment where all prerequisites are properly installed (mainly MarianNMT, sacrebleu, comet-scorer, and SentencePiece).

4.2 External Models

Comparing OPUS-MT to other public models is important to monitor their performance in relation to the state of the art. The extension of the OPUS-MT dashboard to external models is currently supported by the model hub from Hugingface.[12] Metadata tags allow us to search the hub efficiently by filtering by task and language.

		Models	%
(1)	Text2TextGeneration/Translation	11,662	100.00
(2)	Helsinki-NLP	1,439	12.34
	Potential candidates	10,223	87.66
(3)	With no language metadata	7,643	65.54
	With only one language tag	2,166	18.57
	With at least two language tags	414	3.55
(4)	Identifiable language direction	144	1.23

Table 2: Statistics of available models on the Huggingface repository with metadata on the targeted tasks.

We proceed as follows. (1) First, we search the hub to select models that are tagged for tasks *Translation* or *Text2TextGeneration*. (2) Then, we discard Helsinki-NLP models which are already included in the dashboard and (3) keep those models that have at least two language tags. Since the platform does not provide source and target tags, (4) we try to infer the language direction from the model's name by using regular expressions, a naive but effective solution. (5) Finally, we only keep the models that can be used with the translation pipeline straight out of the box. During this process, we encountered several issues such as the need for non-standard language parameters (e.g `en_XX` instead of `en`) or the need for batch size optimization.

We report statistics on the models encountered at each of the steps in Table 2. Surprisingly, only 3.55% of the models have at least two language tags and thus, are potential candidates for our purpose. Although we acknowledge that this is a small subset when compared to the extensive range of available models, the scalability of the method is granted as long as there is sufficient metadata. The lack of documentation on the hub is an issue far beyond our reach. However, we advocate for developers to document models to the fullest extent possible.

Furthermore, apart from the models obtained with the process mentioned above, we specially target large multilingual models to cover as many languages as possible. We added the three following models in various sizes: M2M-100 (Fan et al., 2020), No Language Left Behind (NLLB) (NLLB Team et al., 2022b) and MBART (Tang et al., 2020).

[9] https://github.com/Helsinki-NLP/Opus-MT/

[10] https://github.com/Helsinki-NLP/Tatoeba-Challenge/

[11] https://www.csc.fi/

[12] https://huggingface.co/models

4.3 User-Contributed Translations

In addition to incorporating internal and external models with our own evaluations, the dashboard also provides an opportunity for the community to contribute their collected translations. Translations for a specific benchmark can be easily added following a makefile recipe.

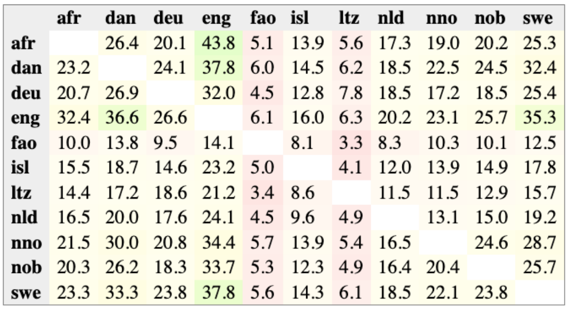
In this context, we have envisioned mainly two scenarios. This feature is highly beneficial to report results for very large MT models that entail high computational costs, such as NLLB's largest variant with 54.5 B parameters. We have collected its scores from `https://tinyurl.com/nllbflorestranslations` and they can be consulted from the dashboard. Secondly, this feature offers flexibility and is especially suitable when researchers aim to include their own models in the dashboard, even if we do not internally run those models. In both cases, scores of user-contributed translations are displayed separately, as their reliability is lower since we did not produce the translations ourselves.

5 MT Performance Dashboard

For the dashboard web-interface, we emphasize a lightweight implementation. In our system, we want to avoid a complex backend and heavy frontends and rather focus on lean and responsive functionalities. The interface has minimal requirements and basically runs with a standard PHP-enabled web server. Data is automatically fetched from the OPUS-MT storage, GitHub or the local file system. Deployment requires no further installation or database setups. The frontend uses cookies and session variables to speed up the process but can also run without them. Server-side caching is used to enable fast response time and no heavy graphics or animations are used that would slow down data transfer and client-side website rendering.

5.1 Benchmark Summaries

The main functionality of the dashboard is to provide summaries of translation performance coming from automatic evaluation. It automatically connects with the relevant repositories described above making their content immediately visible in the interface. Three basic benchmark views are implemented: (i) A summary over best-performing models for a selected language pair over all available benchmarks, (ii) an overview of translation models evaluated on a specific benchmark, and (iii) a comparison of two selected models on available benchmarks (see Figure 1). All views come with simple bar charts and tables linked to relevant information and downloads (see screenshots in the appendix). For multilingual translation models we also added a matrix view in terms of a heatmap illustrating scores across all evaluated language pairs (see Figure 3).



	afr	dan	deu	eng	fao	isl	ltz	nld	nno	nob	swe
afr		26.4	20.1	43.8	5.1	13.9	5.6	17.3	19.0	20.2	25.3
dan	23.2		24.1	37.8	6.0	14.5	6.2	18.5	22.5	24.5	32.4
deu	20.7	26.9		32.0	4.5	12.8	7.8	18.5	17.2	18.5	25.4
eng	32.4	36.6	26.6		6.1	16.0	6.3	20.2	23.1	25.7	35.3
fao	10.0	13.8	9.5	14.1		8.1	3.3	8.3	10.3	10.1	12.5
isl	15.5	18.7	14.6	23.2	5.0		4.1	12.0	13.9	14.9	17.8
ltz	14.4	17.2	18.6	21.2	3.4	8.6		11.5	11.5	12.9	15.7
nld	16.5	20.0	17.6	24.1	4.5	9.6	4.9		13.1	15.0	19.2
nno	21.5	30.0	20.8	34.4	5.7	13.9	5.4	16.5		24.6	28.7
nob	20.3	26.2	18.3	33.7	5.3	12.3	4.9	16.4	20.4		25.7
swe	23.3	33.3	23.8	37.8	5.6	14.3	6.1	18.5	22.1	23.8	

Figure 3: A heatmap of BLEU scores for a multilingual OPUS-MT model covering Germanic languages. Target languages refer to columns in the table.

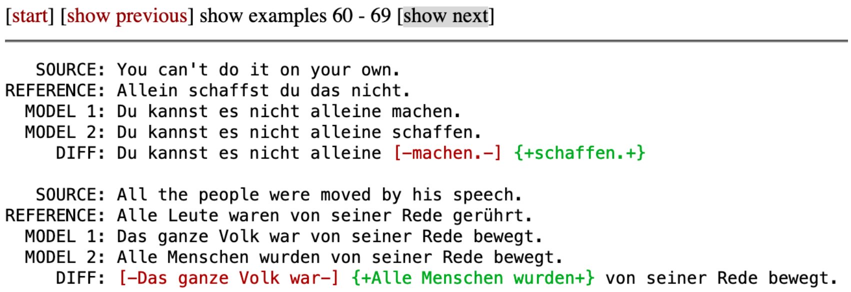
In all modes, the evaluation metric can be selected and other views are linked to quickly jump between them. We provide download links for internal models and links to models' websites when available for metadata regarding each model's characteristics. All system translations and evaluation log files are also available for download to make the process as transparent as possible.

5.2 Inspecting Translations

Another important feature is the possibility to browse through the actual translations produced by each model. We provide all of the translations together with reference translations from the original benchmark in order to study the differences between proposed and human translations. In addition, it is also possible to compare the output of two models on the same benchmark. Highlighting differences can be enabled using a word-level diff function (see Figure 4). The interface displays 10 translation instances at a time but the whole dataset can be downloaded and inspected offline.

6 Related Work

In the current scenario of NLP, which is characterized by the increasing number of available models versus the constant lack of systematic documentation, von Werra et al. (2022) identify three challenges: (1) *reproducibility*, the replicability of model performance, (2) *centralization* to avoid the



[start] [show previous] show examples 60 - 69 [show next]

```
   SOURCE: You can't do it on your own.
REFERENCE: Allein schaffst du das nicht.
  MODEL 1: Du kannst es nicht alleine machen.
  MODEL 2: Du kannst es nicht alleine schaffen.
     DIFF: Du kannst es nicht alleine [-machen.-] {+schaffen.+}

   SOURCE: All the people were moved by his speech.
REFERENCE: Alle Leute waren von seiner Rede gerührt.
  MODEL 1: Das ganze Volk war von seiner Rede bewegt.
  MODEL 2: Alle Menschen wurden von seiner Rede bewegt.
     DIFF: [-Das ganze Volk war-] {+Alle Menschen wurden+} von seiner Rede bewegt.
```

Figure 4: Browsing through translation examples, comparing the output of two models and reference translations for a specific benchmark (Tatoeba English-German).

duplication of workload and one single point of reference, and (3) *coverage*, the inclusion of a diverse set of metrics and languages, as well as pointers for efficiency, bias and robustness.

One of the most well-known efforts on reproducibility is Papers With Code[13] (PWC), an open-source web-based platform that links papers with their implementations and includes many features, such as benchmark evaluations and information on the use of dataset. The Huggingface repository also aims at solving the issue of reproducibility with their `Evaluation on the Hub` (von Werra et al., 2022), a user-friendly platform that allows large-scale evaluation of any of their publicly available models. This feature holds significant potential, however, two primary concerns arise. Firstly, the evaluation is not done systematically, as assessments are performed solely upon user request. Secondly, although the openness of the hub to all users is a move towards democratizing ML, it also results in an absence of metadata for the uploaded resources that make it difficult to find suitable models for every dataset. Yet another similar tool is Dynabench (Kiela et al., 2021), a research platform for dataset creation and dynamic model benchmarking for a wide range of NLP tasks.

Generic platforms mentioned above are useful but make it difficult to get a comprehensive overview of one specific task. MT has a long tradition of shared tasks and several systems have been developed to visualize and analyze benchmark results. The WMT matrix[14] was a dedicated platform for submitting and archiving submissions to the translation tasks at WMT. It provided a useful overview of the state-of-the-art in those tasks, but the system is down.

`MT-CompareEval` (Klejch et al., 2015) partially replaces that service with its WMT instance.[15] The system itself provides a modern tool for the analysis of MT output with various plots of automatic evaluation metrics and an interface for comparing translations with highlighted differences. The software is open source and can be deployed with the option to upload additional system translations, which is also used by the developers to host their own experimental results.[16] In contrast to the OPUS-MT dashboard, it does not implement the replication of translations to verify provided results.

Another open-source tool, `compare-mt` (Neubig et al., 2019), has been developed to explore translation outputs. It supports a deeper analysis and comparison using detailed statistics of word-level and sentence-level accuracies, salient n-grams, etc.

On the commercial side, there is Intento, a language service provider that publishes a yearly report with an overview of current MT systems Savenkov and Lopez (2022) with a focus on commercial models. They provide an end-to-end MT evaluation service that comes with a cost and the yearly evaluation is not transparent and open.

7 Conclusions and Future Work

The OPUS-MT dashboard implements a simple yet comprehensive interface for a systematic evaluation of public translation models. The main purpose is to provide an overview of OPUS-MT models and to relate their performance to other openly available models. The focus is on verifiable performance and a centralized evaluation procedure. The workflow and collection stress transparency and replicability and can easily be extended with new models and benchmarks.

The current implementation is fully functional

[13] https://paperswithcode.com

[14] http://matrix.statmt.org

[15] http://wmt.ufal.cz/

[16] http://mt-compareval.ufal.cz/

Provider	Model	Parameter Size	Batch size	Benchmarks / hour
facebook	m2m100_418M	418M	16	17.47
facebook	m2m100_418M	418M	16	14.14
facebook	m2m100_418M	418M	16	15.25
facebook	m2m100_1.2B	1.2B	8	8.97
facebook	m2m100_1.2B	1.2B	8	8.14
facebook	m2m100_1.2B	1.2B	8	10.47
facebook	nllb3.3B	3.3B	16	11.03
facebook	nllb3.3B	3.3B	16	11.78
facebook	nllb3.3B	3.3B	16	12.03
facebook	nllb3.3B	3.3B	2	4.18
facebook	nllb3.3B	3.3B	2	3.94
facebook	nllb3.3B	3.3B	2	3.19

Table 3: Approximate inference statistics (number of translated benchmarks per GPU hour) on three large multilingual models with different sizes and different runs on a single NVIDIA Ampere A100 GPU.

but we already work on several extensions. First of all, we would like to integrate more information about the model properties in the dashboard. Important features are model size, inference time and computational costs that can be related to translation performance. Additionally, we want to tag other important characteristics such as multilingual versus bilingual models. Heatmaps for comparing multilingual model scores are also on our to-do list as well as better overviews of top-scores in multilingual benchmarks. We also want to integrate our geolocated visualization of language coverage implemented in OPUS-MT map.[17]

Finally, we are continuously working on the integration of new benchmarks and the systematic evaluation of available models. We look into other released models and their use for replicating benchmark results. We also continue to collect benchmarks and will integrate sentence-level scores while browsing through translation output. We may also connect to other systems like `MT-CompareEval` for more detailed analyses.

Limitations

In this paper, we have introduced OPUS-MT dashboard, our system for MT evaluation with a focus on centralization and reproducibility. One of the limitations of the presented approach is that the current coverage is based solely on automatic MT metrics. Nevertheless, as mentioned above, we are working towards adding pointers for model size, inference time and computational costs. A large scale manual assessment is beyond our capabilities. However, we consider the option to enable community-driven feedback that could help to add human judgments to the system outputs. Furthermore, at the moment we are not aware of how we can add information regarding bias and fairness, but we will look into additional points of information that can be added to the collection.

Another limitation of our method is that for the multilingual external models, currently we only provide English-centric translations (en-xx, xx-en), due to the high computational costs of running inference on large language and translation models as shown in Table 3. We will incrementally close the gaps and maintain a systematic approach to update the dashboard. We also hope that the dashboard will trigger more models to become available and that metadata for their re-use will be improved.

Finally, the current implementation is limited to single-reference benchmarks and the pipelines assume sentence-level translation. However, multi-reference test sets are extremely rare but we will still consider a support of such data sets in the future. Document-level translation will be important in the near future and for that we will need to adjust our workflow.

Broader Impacts

The OPUS-MT dashboard has the potential to significantly impact the field of MT research by providing a centralized tool for visualizing and evaluating the quality of MT output in a systematic manner. We hope for it to become a point of reference where everyone (1) can consult which model suits best their use case by answering *"Which model should I use for language pair X and domain Y?"* and (2) can obtain proper baselines during paper writing without the need to run again the same experiments, saving time and, more importantly, computational costs. We provide selected rough figures of in-

[17] `https://github.com/Helsinki-NLP/OPUS-MT-map`

ference speed in terms of translated benchmarks per GPU hour in Table 3 to illustrate the carbon footprint generated by running the models.

Furthermore, we hope that the overall picture that the OPUS-MT dashboard offers on MT for specific language pairs will encourage the development of resources for low-resource language pairs making it possible to see where there are gaps or where multilingual models fail to deliver.

Acknowledgements

This work was supported by the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 771113.

This work was also supported by the HPLT project which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070350.

References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, et al. 2020. Tico-19: the translation initiative for covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Anastasopoulos Antonios, Barrault Loc, Luisa Bentivogli, Marcely Zanon Boito, Bojar Ondřej, Roldano Cattoni, Currey Anna, Dinu Georgiana, Duh Kevin, Elbayad Maha, et al. 2022. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

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.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. Ntrex-128–news test references for mt evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Ondrej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. Mt-compareval: Graphical evaluation interface for machine translation development. *Prague Bull. Math. Linguistics*, 104:63–74.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, and John Wieting. 2019. compare-mt: A tool for holistic comparison of language generation systems. *CoRR*, abs/1903.07926.

Tommi Nieminen. 2021. OPUS-CAT: Desktop NMT with CAT integration and local fine-tuning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 288–294, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022a. No language left behind: Scaling human-centered machine translation.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022b. No language left behind: Scaling human-centered machine translation.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Konstantin Savenkov and Michel Lopez. 2022. The state of the machine translation 2022. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 32–49, Orlando, USA. Association for Machine Translation in the Americas.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Jörg Tiedemann. 2020. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182.

Jörg Tiedemann, Mikko Aulamo, Sam Hardwick, and Tommi Nieminen. 2023. *Open Translation Models, Tools and Services*, pages 325–330. Springer International Publishing, Cham.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2022. Democratizing machine translation with opus-mt.

Leandro von Werra, Lewis Tunstall, Abhishek Thakur, Alexandra Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, Helen Ngo, et al. 2022. Evaluate & evaluation on the hub: Better best practices for data and model measurement. *arXiv preprint arXiv:2210.01970*.

Andy B Yoo, Morris A Jette, and Mark Grondona. 2003. Slurm: Simple linux utility for resource management. In *Job Scheduling Strategies for Parallel Processing: 9th International Workshop, JSSPP 2003, Seattle, WA, USA, June 24, 2003. Revised Paper 9*, pages 44–60. Springer.

A Evaluation Hyperparameters

Hyperparameter	Value
beam-size	4
max-length	500
max-length-crop	True
maxi-batch	512
maxi-batch-sort	src
mini-batch	256

Table 4: Marian hyperparameters for decoding internal models.

Hyperparameter	Value
batch_size	64
do_sample	False
max_length	500
num_beams	1
top_k	50

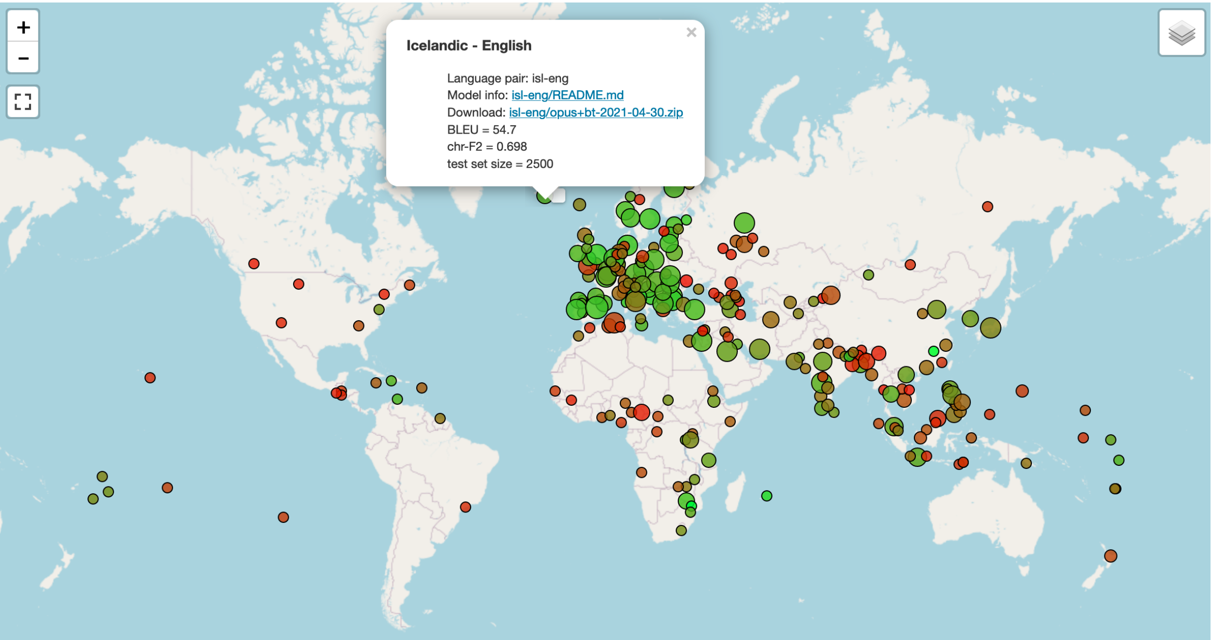
Table 5: Hyperparameters for decoding external models with HuggingFace's translation pipeline.

B OPUS-MT Map

The OPUS-MT map is yet another visualization of the availability of machine translation models focusing on the language coverage in some geographic distribution. The main purpose is to illustrate the concentration of work on specific regions without making strong claims about the location of specific languages and language speakers. Geolocations are taken from Glottolog and the interactive map supports the visualization of translation for language pairs in both directions using a number of selected benchmarks like the Tatoeba translation challenge and the Flores benchmark. A screenshot of the map is shown in Figure 5.

C Dashboard Screenshots

This appendix shows a number of screenshots from the live dashboard at `https://opus.nlpl.eu/dashboard`. We include different variants of performance plots and show only BLEU-score-based evaluations in the examples shown below. The dashboard makes it possible to show the best performing models for a specific language pair across all benchmarks (see Figure 7), the performance of all models evaluated for a specific benchmark



Icelandic - English

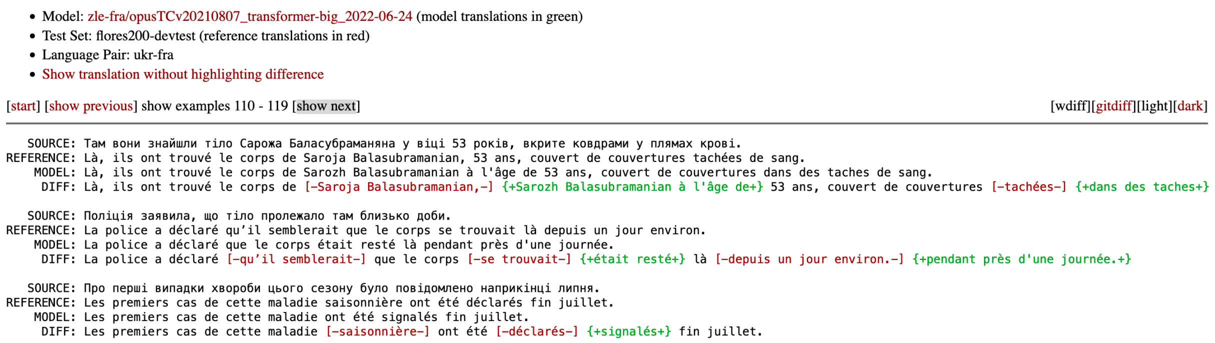
Language pair: isl-eng
Model info: isl-eng/README.md
Download: isl-eng/opus+bt-2021-04-30.zip
BLEU = 54.7
chr-F2 = 0.698
test set size = 2500

Figure 5: A geographic visualization of the language coverage in public OPUS-MT models. The map plots models according to Tatoeba translation challenge results indicating performance based on color (green is good and red is low performance in BLEU). Smaller circles indicate smaller test sets and are, therefore, less reliable.

(Figure 8), an overview of benchmark results for a selected model including multilingual models (Figure 9) and a comparison of benchmark results for two selected models (Figure 10). An example of the translation inspection feature is shown in Figure 6.

D Video Demonstration

A video demonstration of the system can be accessed at `https://youtu.be/K2cKoAt3AIY`.



- Model: zle-fra/opusTCv20210807_transformer-big_2022-06-24 (model translations in green)
- Test Set: flores200-devtest (reference translations in red)
- Language Pair: ukr-fra
- Show translation without highlighting difference

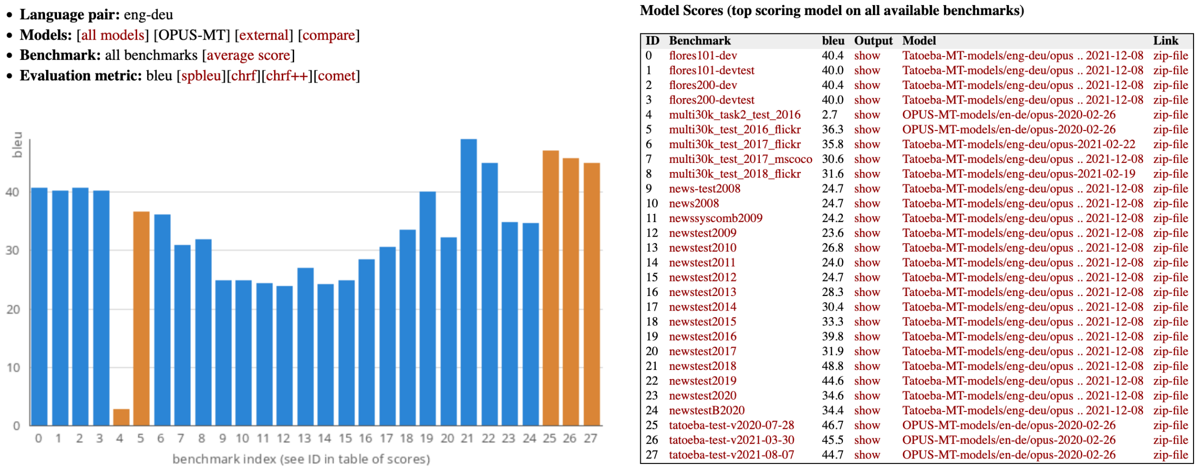
[start] [show previous] show examples 110 - 119 [show next] [wdiff][gitdiff][light][dark]

```
   SOURCE: Там вони знайшли тіло Сарожа Баласубраманяна у віці 53 років, вкрите ковдрами у плямах крові.
REFERENCE: Là, ils ont trouvé le corps de Saroja Balasubramanian, 53 ans, couvert de couvertures tachées de sang.
    MODEL: Là, ils ont trouvé le corps de Sarozh Balasubramanian à l'âge de 53 ans, couvert de couvertures dans des taches de sang.
     DIFF: Là, ils ont trouvé le corps de [-Saroja Balasubramanian,-] {+Sarozh Balasubramanian à l'âge de+} 53 ans, couvert de couvertures [-tachées-] {+dans des taches+}

   SOURCE: Поліція заявила, що тіло пролежало там близько доби.
REFERENCE: La police a déclaré qu’il semblerait que le corps se trouvait là depuis un jour environ.
    MODEL: La police a déclaré que le corps était resté là pendant près d'une journée.
     DIFF: La police a déclaré [-qu’il semblerait-] que le corps [-se trouvait-] {+était resté+} là [-depuis un jour environ.-] {+pendant près d'une journée.+}

   SOURCE: Про перші випадки хвороби цього сезону було повідомлено наприкінці липня.
REFERENCE: Les premiers cas de cette maladie saisonnière ont été déclarés fin juillet.
    MODEL: Les premiers cas de cette maladie ont été signalés fin juillet.
     DIFF: Les premiers cas de cette maladie [-saisonnière-] ont été [-déclarés-] {+signalés+} fin juillet.
```

Figure 6: Browsing through benchmark translations with highlighting differences between reference and system translation. Here a sample from the Flores200 devtest set for Ukrainian to French translated by an OPUS-MT model for East Slavic languages to French.

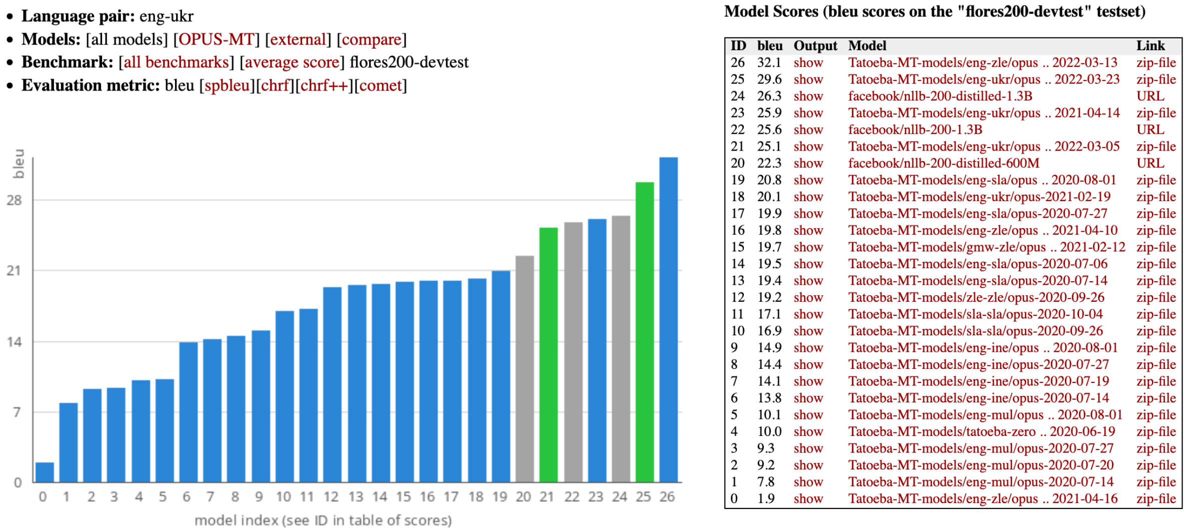


- **Language pair:** eng-deu
- **Models:** [all models] [OPUS-MT] [external] [compare]
- **Benchmark:** all benchmarks [average score]
- **Evaluation metric:** bleu [spbleu][chrf][chrf++][comet]

Model Scores (top scoring model on all available benchmarks)

ID	Benchmark	bleu	Output	Model	Link
0	flores101-dev	40.4	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
1	flores101-devtest	40.0	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
2	flores200-dev	40.4	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
3	flores200-devtest	40.0	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
4	multi30k_task2_test_2016	2.7	show	OPUS-MT-models/en-de/opus-2020-02-26	zip-file
5	multi30k_test_2016_flickr	36.3	show	OPUS-MT-models/en-de/opus-2020-02-26	zip-file
6	multi30k_test_2017_flickr	35.8	show	Tatoeba-MT-models/eng-deu/opus-2021-02-22	zip-file
7	multi30k_test_2017_mscoco	30.6	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
8	multi30k_test_2018_flickr	31.6	show	Tatoeba-MT-models/eng-deu/opus-2021-02-19	zip-file
9	news-test2008	24.7	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
10	news2008	24.7	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
11	newssyscomb2009	24.2	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
12	newstest2009	23.6	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
13	newstest2010	26.8	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
14	newstest2011	24.0	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
15	newstest2012	24.7	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
16	newstest2013	28.3	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
17	newstest2014	30.4	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
18	newstest2015	33.3	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
19	newstest2016	39.8	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
20	newstest2017	31.9	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
21	newstest2018	48.8	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
22	newstest2019	44.6	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
23	newstest2020	34.6	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
24	newstestB2020	34.4	show	Tatoeba-MT-models/eng-deu/opus .. 2021-12-08	zip-file
25	tatoeba-test-v2020-07-28	46.7	show	OPUS-MT-models/en-de/opus-2020-02-26	zip-file
26	tatoeba-test-v2021-03-30	45.5	show	OPUS-MT-models/en-de/opus-2020-02-26	zip-file
27	tatoeba-test-v2021-08-07	44.7	show	OPUS-MT-models/en-de/opus-2020-02-26	zip-file

Figure 7: Best performing OPUS-MT models on English-German benchmarks.

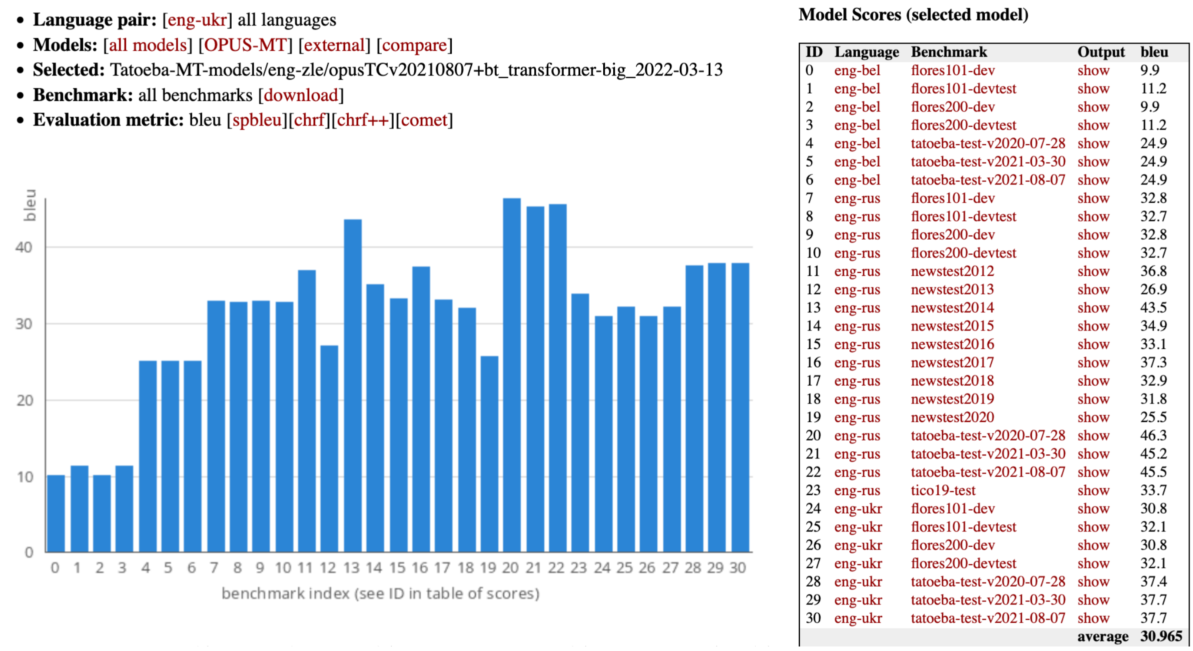


- **Language pair:** eng-ukr
- **Models:** [all models] [OPUS-MT] [external] [compare]
- **Benchmark:** [all benchmarks] [average score] flores200-devtest
- **Evaluation metric:** bleu [spbleu][chrf][chrf++][comet]

Model Scores (bleu scores on the "flores200-devtest" testset)

ID	bleu	Output	Model	Link
26	32.1	show	Tatoeba-MT-models/eng-zle/opus .. 2022-03-13	zip-file
25	29.6	show	Tatoeba-MT-models/eng-ukr/opus .. 2022-03-23	zip-file
24	26.3	show	facebook/nllb-200-distilled-1.3B	URL
23	25.9	show	Tatoeba-MT-models/eng-ukr/opus .. 2021-04-14	zip-file
22	25.6	show	facebook/nllb-200-1.3B	URL
21	25.1	show	Tatoeba-MT-models/eng-ukr/opus .. 2022-03-05	zip-file
20	22.3	show	facebook/nllb-200-distilled-600M	URL
19	20.8	show	Tatoeba-MT-models/eng-sla/opus .. 2020-08-01	zip-file
18	20.1	show	Tatoeba-MT-models/eng-ukr/opus-2021-02-19	zip-file
17	19.9	show	Tatoeba-MT-models/eng-sla/opus-2020-07-27	zip-file
16	19.8	show	Tatoeba-MT-models/eng-zle/opus .. 2021-04-10	zip-file
15	19.7	show	Tatoeba-MT-models/gmw-zle/opus .. 2021-02-12	zip-file
14	19.5	show	Tatoeba-MT-models/eng-sla/opus-2020-07-06	zip-file
13	19.4	show	Tatoeba-MT-models/eng-sla/opus-2020-07-14	zip-file
12	19.2	show	Tatoeba-MT-models/zle-zle/opus-2020-09-26	zip-file
11	17.1	show	Tatoeba-MT-models/sla-sla/opus-2020-10-04	zip-file
10	16.9	show	Tatoeba-MT-models/sla-sla/opus-2020-09-26	zip-file
9	14.9	show	Tatoeba-MT-models/eng-ine/opus .. 2020-08-01	zip-file
8	14.4	show	Tatoeba-MT-models/eng-ine/opus-2020-07-27	zip-file
7	14.1	show	Tatoeba-MT-models/eng-ine/opus-2020-07-19	zip-file
6	13.8	show	Tatoeba-MT-models/eng-ine/opus-2020-07-14	zip-file
5	10.1	show	Tatoeba-MT-models/eng-mul/opus .. 2020-08-01	zip-file
4	10.0	show	Tatoeba-MT-models/tatoeba-zero .. 2020-06-19	zip-file
3	9.3	show	Tatoeba-MT-models/eng-mul/opus-2020-07-27	zip-file
2	9.2	show	Tatoeba-MT-models/eng-mul/opus-2020-07-20	zip-file
1	7.8	show	Tatoeba-MT-models/eng-mul/opus-2020-07-14	zip-file
0	1.9	show	Tatoeba-MT-models/eng-zle/opus .. 2021-04-16	zip-file

Figure 8: List of models that support translating from English to Ukrainian. Blue bars refer to OPUS-MT models, green bars are compact student models and grey bars refer to external models.

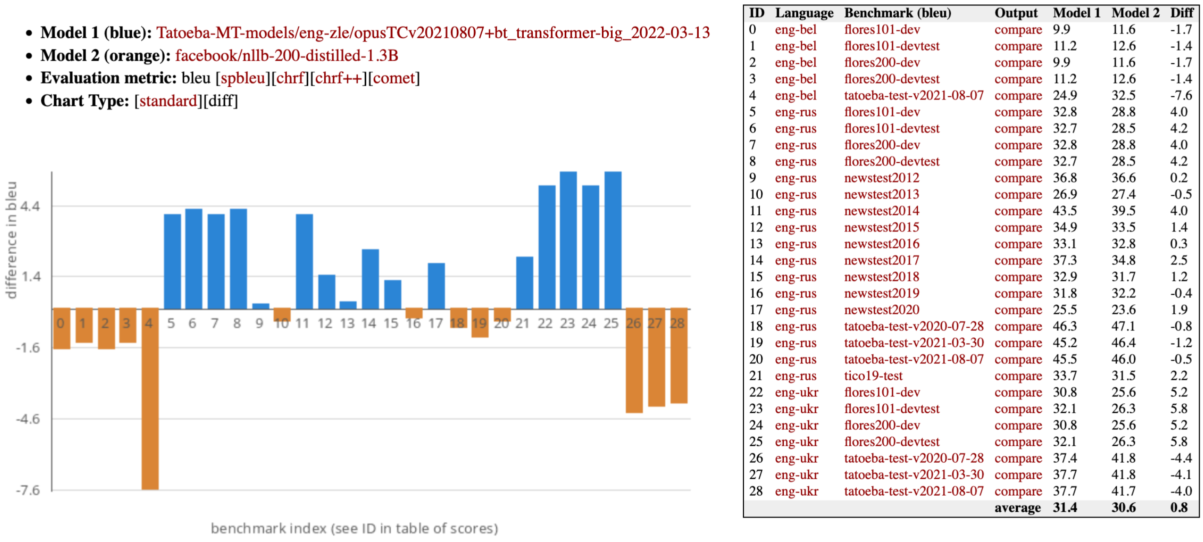


- **Language pair:** [eng-ukr] all languages
- **Models:** [all models] [OPUS-MT] [external] [compare]
- **Selected:** Tatoeba-MT-models/eng-zle/opusTCv20210807+bt_transformer-big_2022-03-13
- **Benchmark:** all benchmarks [download]
- **Evaluation metric:** bleu [spbleu][chrf][chrf++][comet]

Model Scores (selected model)

ID	Language	Benchmark	Output	bleu
0	eng-bel	flores101-dev	show	9.9
1	eng-bel	flores101-devtest	show	11.2
2	eng-bel	flores200-dev	show	9.9
3	eng-bel	flores200-devtest	show	11.2
4	eng-bel	tatoeba-test-v2020-07-28	show	24.9
5	eng-bel	tatoeba-test-v2021-03-30	show	24.9
6	eng-bel	tatoeba-test-v2021-08-07	show	24.9
7	eng-rus	flores101-dev	show	32.8
8	eng-rus	flores101-devtest	show	32.7
9	eng-rus	flores200-dev	show	32.8
10	eng-rus	flores200-devtest	show	32.7
11	eng-rus	newstest2012	show	36.8
12	eng-rus	newstest2013	show	26.9
13	eng-rus	newstest2014	show	43.5
14	eng-rus	newstest2015	show	34.9
15	eng-rus	newstest2016	show	33.1
16	eng-rus	newstest2017	show	37.3
17	eng-rus	newstest2018	show	32.9
18	eng-rus	newstest2019	show	31.8
19	eng-rus	newstest2020	show	25.5
20	eng-rus	tatoeba-test-v2020-07-28	show	46.3
21	eng-rus	tatoeba-test-v2021-03-30	show	45.2
22	eng-rus	tatoeba-test-v2021-08-07	show	45.5
23	eng-rus	tico19-test	show	33.7
24	eng-ukr	flores101-dev	show	30.8
25	eng-ukr	flores101-devtest	show	32.1
26	eng-ukr	flores200-dev	show	30.8
27	eng-ukr	flores200-devtest	show	32.1
28	eng-ukr	tatoeba-test-v2020-07-28	show	37.4
29	eng-ukr	tatoeba-test-v2021-03-30	show	37.7
30	eng-ukr	tatoeba-test-v2021-08-07	show	37.7
			average	**30.965**

Figure 9: Benchmark results for a multilingual model translating English to East Slavic languages.



- **Model 1 (blue):** Tatoeba-MT-models/eng-zle/opusTCv20210807+bt_transformer-big_2022-03-13
- **Model 2 (orange):** facebook/nllb-200-distilled-1.3B
- **Evaluation metric:** bleu [spbleu][chrf][chrf++][comet]
- **Chart Type:** [standard][diff]

ID	Language	Benchmark (bleu)	Output	Model 1	Model 2	Diff
0	eng-bel	flores101-dev	compare	9.9	11.6	-1.7
1	eng-bel	flores101-devtest	compare	11.2	12.6	-1.4
2	eng-bel	flores200-dev	compare	9.9	11.6	-1.7
3	eng-bel	flores200-devtest	compare	11.2	12.6	-1.4
4	eng-bel	tatoeba-test-v2021-08-07	compare	24.9	32.5	-7.6
5	eng-rus	flores101-dev	compare	32.8	28.8	4.0
6	eng-rus	flores101-devtest	compare	32.7	28.5	4.2
7	eng-rus	flores200-dev	compare	32.8	28.8	4.0
8	eng-rus	flores200-devtest	compare	32.7	28.5	4.2
9	eng-rus	newstest2012	compare	36.8	36.6	0.2
10	eng-rus	newstest2013	compare	26.9	27.4	-0.5
11	eng-rus	newstest2014	compare	43.5	39.5	4.0
12	eng-rus	newstest2015	compare	34.9	33.5	1.4
13	eng-rus	newstest2016	compare	33.1	32.8	0.3
14	eng-rus	newstest2017	compare	37.3	34.8	2.5
15	eng-rus	newstest2018	compare	32.9	31.7	1.2
16	eng-rus	newstest2019	compare	31.8	32.2	-0.4
17	eng-rus	newstest2020	compare	25.5	23.6	1.9
18	eng-rus	tatoeba-test-v2020-07-28	compare	46.3	47.1	-0.8
19	eng-rus	tatoeba-test-v2021-03-30	compare	45.2	46.4	-1.2
20	eng-rus	tatoeba-test-v2021-08-07	compare	45.5	46.0	-0.5
21	eng-rus	tico19-test	compare	33.7	31.5	2.2
22	eng-ukr	flores101-dev	compare	30.8	25.6	5.2
23	eng-ukr	flores101-devtest	compare	32.1	26.3	5.8
24	eng-ukr	flores200-dev	compare	30.8	25.6	5.2
25	eng-ukr	flores200-devtest	compare	32.1	26.3	5.8
26	eng-ukr	tatoeba-test-v2020-07-28	compare	37.4	41.8	-4.4
27	eng-ukr	tatoeba-test-v2021-03-30	compare	37.7	41.8	-4.1
28	eng-ukr	tatoeba-test-v2021-08-07	compare	37.7	41.7	-4.0
			average	**31.4**	**30.6**	**0.8**

Figure 10: Comparison of benchmark results between an OPUS-MT model for English to East Slavic languages and the distilled NLLB model with 1.3B parameters.