MT-Lens: An all-in-one Toolkit for Better Machine Translation Evaluation

Javier García Gilabert, Carlos Escolano, Audrey Mash, Xixian Liao, Maite Melero
Barcelona Super Computing Center (BSC)
{javier.garcia1,carlos.escolano,
audrey.mash,xixian.liao,maite.melero}@bsc.es

Abstract

We introduce MT-Lens[1], a framework designed to evaluate Machine Translation (MT) systems across a variety of tasks, including translation quality, gender bias detection, added toxicity, and robustness to misspellings. While several toolkits have become very popular for benchmarking the capabilities of Large Language Models (LLMs), existing evaluation tools often lack the ability to thoroughly assess the diverse aspects of MT performance. MT-Lens addresses these limitations by extending the capabilities of LM-eval-harness for MT, supporting state-of-the-art datasets and a wide range of evaluation metrics. It also offers a user-friendly platform to compare systems and analyze translations with interactive visualizations. MT-Lens aims to broaden access to evaluation strategies that go beyond traditional translation quality evaluation, enabling researchers and engineers to better understand the performance of a NMT model and also easily measure system's biases.

1 Introduction

Neural Machine Translation (NMT) models are typically evaluated using automated metrics that compare the model's translated outputs to one or more reference translations. Recent neural-based evaluation metrics have demonstrated high correlations with human judgments (Rei et al., 2022a; Sellam et al., 2020; Juraska et al., 2023) replacing canonical overlap-based metrics (Popović, 2015; Papineni et al., 2002). While automated metrics provide an essential means for assessing quality improvements and have become the way to go in evaluating state-of-the-art NMT systems, they only offer a general intuition of the model's overall performance and often lack interpretability (Perrella et al., 2024). As such, the reliance on these metrics raises concerns about their specific error types that might affect the end-user experience, such as reinforcing gender bias (Zaranis et al., 2024), translationese or language mismatch (Zouhar et al., 2024).

To address these problems, recent works have focused on developing inherently interpretable metrics for MT evaluation (Guerreiro et al., 2023b) via token-level annotations that offer a more granular-insight at the segment level. However, such limitations necessitate more granular evaluation tools that can dissect translation outputs to identify and analyze specific error types and their impact on overall quality, thereby enabling MT engineers to make more informed decisions when evaluating translation systems.

Moreover, existing evaluation methodologies in state-of-the-art NMT models primarily focus on evaluating translation quality, frequently overlooking other equally critical evaluations like gender bias, added toxicity or robustness to misspellings. These biases and harmful outputs can have significant consequences for users (Savoldi et al., 2024a), highlighting a need for evaluation strategies that go beyond quality to encompass a broader range of tasks.

MT-Lens seeks to address these critical gaps by providing a unified framework to test generative language models on a number of different machine translation evaluation tasks and providing a user-friendly interface to analyze the results. MT-Lens is based on LM-eval-harness (Gao et al., 2024) which has been widely used for evaluating LLMs in several Natural Language Understanding (NLU) tasks. Building upon this comprehensive framework, MT-Lens extends the evaluation capabilities of LM-eval-harness for NMT. Contributions of this framework are listed as follows:

- We support novel evaluation tasks for detecting gender bias, added toxicity, and robustness to character noise in MT.

[1] We release our code at https://github.com/langtech-bsc/mt-evaluation. Demo is available at this link while the demo video is available at this link.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 51–60
April 30, 2025 ©2025 Association for Computational Linguistics

Task	Dataset	Task name	Languages
General-MT	FLORES-200 (Costa-jussà et al., 2022)	{src}_{tgt}_flores_{split}	200
	NTREX-128 (Federmann et al., 2022)	{src}_{tgt}_ntrex	128
	TATOEBA (Tiedemann, 2020)	{src}_{tgt}_tatoeba	555
	NTEU (Bié et al., 2020)	{src}_{tgt}_nteu	25
Added Toxicity	HOLISTICBIAS (Smith et al., 2022)	{src}_{tgt}_{axis}_hb	1
Gender Bias	MUST-SHE (Bentivogli et al., 2020)	{src}_{tgt}_must_she	5
	MMHB (Tan et al., 2024)	{src}_{tgt}_mmhb_{split}	7
	MT-GENEVAL (Currey et al., 2022)	{src}_{tgt}_geneval_{split}	8
Robustness to Character Noise	FLORES-200 devtest	{src}_{tgt}_perturbations	200

Table 1: Machine translation datasets natively supported by MT-LENS grouped by task type.

- We support a variety of state-of-the-art benchmark datasets and evaluation metrics for assessing translation quality.
- We provide a user-friendly interface with interactive visualizations at both segment and system levels, enabling thorough error analysis and performance assessment of evaluations performed using MT-LENS.
- We provide bootstrapping significance tests for comparing MT systems on both neural-based and overlap-based machine translation metrics.

MT-LENS is designed to broaden access to novel evaluation tasks that go beyond traditional translation quality in MT, while also supporting general NLU tasks. The framework is maintained by the Language Technologies Unit at the Barcelona Supercomputing Center, ensuring ongoing updates with the latest features of LM-eval-harness and support for new MT datasets developed by the research community.

2 Related work

Typically, automatic metrics like BLEU (Papineni et al., 2002) are simply applied and reported at the corpus level. In recent years, the evolution of MT evaluation has seen the development of tools that offer more granular insights.

MT-COMPAREEVAL (Klejch et al., 2015) provides comparative analysis of segment-level errors, focusing on identifying differences in n-grams between two MT outputs. Similarly, COMPARE-MT (Neubig et al., 2019) offers a holistic analysis for pairs of MT systems, examining performance metrics such as n-gram frequency and part-of-speech accuracy. MATEO (Vanroy et al., 2023) offers a friendly web-based interface for evaluating MT outputs on several evaluation metrics. MT-TELESCOPE, a newer platform developed by Rei et al. (2021), not only supports segment-level analysis but also offers a web-based interface for better visualization of comparative performance between two MT systems. In addition, it enables focused analysis on phenomena like named entity translation and terminology handling. The last three mentioned tools include statistical significance testing through bootstrapped t-tests to ensure the reliability and statistical validity of their evaluations. Importantly, these tools require users to upload the source sentences and system outputs to perform evaluations.

More recently, TOWEREVAL (Alves et al., 2024) has been developed specifically for LLMs. This evaluation framework allows users to benchmark their models against a comprehensive suite of datasets for evaluating translation quality. It supports both generation and evaluation processes, allowing users to run inference and compute a variety of metrics such as BLEU, COMET (Rei et al., 2022a), COMET-KIWI (Rei et al., 2022b), CHRF (Popović, 2015) and TER (Snover et al., 2006). It also allows for the creation of custom test suites and instructions.

Although the aforementioned works aimed to become the standard tools for evaluating NMT systems, they have not achieved widespread adoption within the MT research community. On the contrary, the LM-eval-harness library has gained significant popularity in the NLP community due to its extensibility, modularity, and ease of use. By building on this widely adopted framework, MT-LENS bridges the gap between the versatility of LM-eval-harness and the specific needs of MT evaluation, thereby addressing the limitations of prior works.

MT-LENS is largely inspired by previous frameworks but differentiates itself by being a unified framework where the user can easily run evaluations on the desired model and visualize the results in a user-friendly interface. It also supports a broader spectrum of MT tasks, including bias detection, toxicity evaluation, and robustness to character noise, in addition to traditional translation quality evaluation.

3 Building blocks

MT-LENS follows a similar evaluation methodology as LM-eval-harness for MT tasks. When evaluating a system, it follows a predefined sequence of steps which can be divided into five main blocks: 🟦 Models, 🟨 Tasks, 🟩 Format, 🟪 Metrics and 🟥 Results. Once the evaluation is completed the user interface will show a fine-grained analysis of the system with interactive visualizations.

🟦 **Models** MT-LENS supports different inference frameworks for running MT tasks: `fairseq` (Ott et al., 2019), `CTranslate2`[2], `transformers` (Wolf et al., 2020), `vllm` (Kwon et al., 2023), and others. If a model is not directly supported, users can utilize the `simplegenerator` wrapper, which accepts pre-generated translations from a text file.

🟨 **Tasks** A MT task is defined by the dataset to be used and the language pair involved (Table 1). Each MT task is uniquely named using the convention:

```
{source language}_{target language}_{dataset}
```

🟩 **Format** When executing MT tasks we need to define how input prompts are formatted for the selected model. Different models may require the source sentence to be formatted in a specific style. Users can specify the desired template through a YAML file, and the task implementation will automatically format the source sentence accordingly.

🟪 **Metrics** MT-LENS includes an extensive number of evaluation metrics for MT tasks. These metrics cover both overlap-based and neural-based metrics which are listed in Table 2. Metrics are computed at the segment level and then aggregated at the system level. Each metric has some configurable hyper-parameters that users can adjust through a YAML file.

🟥 **Results** Evaluation results are outputted in JSON format including source sentences, reference translations, aggregated metrics and segment level scores. Each JSON evaluation file is then used by the MT-LENS UI to provide a more intuitive analysis for the user.

3.1 Example usage

Given an NMT model and a specified task, we can evaluate the model using MT-LENS with the following command:

```
model='./models/madlad400/'
output_dir='results/results.json'

lm_eval --model hf \
  --model_args "pretrained=${model}" \
  --tasks en_ca_flores_devtest \
  --output_path $output_dir \
  --translation_kwargs "
                 src_language=eng_Latn,
                 tgt_language=cat_Latn,
                 prompt_style=madlad400
                 "
```

where the `hf` model argument indicates that the model is implemented using `transformers`. Then, we specify the model path, source and target languages, and prompt style. The task is set to `en_ca_flores_devtest`, and the results will be saved to the specified output directory.

3.2 MT Tasks

In this section we outline the MT related tasks implemented in MT-LENS.

General-MT This task consists in evaluating the faithfulness and the quality of the translation using reference-based and quality estimation metrics. We show in Table 1 the datasets that are natively supported in the MT-LENS framework.

Added toxicity This type of toxicity arises when a toxic element appears in the translated sentence without a corresponding toxic element in the source sentence, or when a toxic element in the translation results from a mistranslation of a non-toxic element in the source sentence. We use the HOLISTICBIAS dataset (Smith et al., 2022) to evaluate NMT models on this task, which has previously been used for identifying added toxicity in NMT models (García Gilabert et al., 2024; Costa-jussà et al., 2024a). HOLISTICBIAS consists of approximately 472k sentences in English that are created using sentence templates across 13 demographic axes (gender, ability, religion,

[2] https://github.com/OpenNMT/CTranslate2

Type	Name	Implementation
Overlap Reference-based	BLEU (Papineni et al., 2002) TER (Snover et al., 2006) ChrF (Popović, 2015)	`SacreBLEU` (Post, 2018) `SacreBLEU` `SacreBLEU`
Neural Reference-based	COMET (Rei et al., 2022a) BLEURT (Sellam et al., 2020) MetricX	`unbabel-comet` (Stewart et al., 2020) `transformers` (Wolf et al., 2020) `metricx` (Juraska et al., 2023)
Neural Reference-based and error span	XComet (Guerreiro et al., 2023b)	`unbabel-comet`
Quality estimation	MetricX-QE Comet-Kiwi (Rei et al., 2022b)	`metricx` `unbabel-comet`
Quality estimation and error span	XComet-QE	`unbabel-comet`
Word lists	ETOX	`nllb` (Costa-jussà et al., 2022)
Embedding-based	MuTox (Costa-jussà et al., 2024b) Detoxify	`seamless` (Barrault et al., 2023) `detoxify` (Hanu and Unitary team, 2020)

Table 2: Evaluation metrics supported by MT-Lens.

etc.). For measuring added toxicity we first filter the source sentences using MuTox (Costa-jussà et al., 2024b) as done in (Tan et al., 2024) and measure the toxicity in the translations using ETOX (Costa-jussà et al., 2023), MuTox (Costa-jussà et al., 2024b) and Detoxify (Hanu and Unitary team, 2020) toxicity classifiers. Since ETOX supports 200 languages, it allows us to evaluate a wide range of languages for added toxicity using English as the source language. We then measure the translation faithfulness using Comet-Kiwi on the toxic sentences detected by each toxicity classifier as it has been proved useful to evaluate hallucinations when no reference is available (Guerreiro et al., 2023a).

Gender bias This type of bias in translation occurs when the system's prediction is skewed toward a specific gender due to stereotypes or inequalities (Friedman and Nissenbaum, 1995; Savoldi et al., 2024b; Sant et al., 2024). Additionally, as not all languages contain the same amount of gender information, when translating from a notionally gendered or ungendered language to a grammatically gendered language, decisions about gender assignation may need to be made from little or no context. This leads to gender bias in cases where gender is consistently assigned following stereotypical patterns, such as labelling all nurses as female and all doctors as male.

We implement three tasks for the evaluation of gender bias when translating out of English, one using the MuST-SHE dataset (Bentivogli et al., 2020; Mash et al., 2024) one using the Massive Multilingual Holistic Bias Dataset (MMHB) (Tan et al., 2024), and finally, one using MT-GenEval (Currey et al., 2022).

MuST-SHE contains approximately 1000 English sentences. All English terms that must be assigned a gender in translation have been identified, and tuples containing both the correctly and incorrectly gendered translated forms are provided for evaluation. The use of transcripts of natural speech allows for the assessment of gender bias in more complex contextual clues and co-reference situations compared to template-based datasets. The accuracy of sex is measured using the revised script of (Mash et al., 2024) and is reported at the sentence and data set level, allowing for a fine-grained analysis of model performance.

MMHB makes use of placeholder-based sentence generation to generate feminine, masculine and neutral variations of each sentence pattern (Tan et al., 2024). The dataset consists of 152,720 English sentences partitioned into train, dev and devtest splits. The placeholder-based system allows for the creation of sentences covering all variations of morphological agreement in the target languages. The created sentences are then organized into groupings and the ChrF scores are measured for the different subsets.

MT-GenEval consists of two distinct tasks, one operating at sentence level and the other looking at accuracy when extracting gender from the preceding context. In both cases, all data has been human-reviewed to exclude sentences with ambiguous gender references, and counterfactual data created. When dealing with contextual inputs, professions are grouped into stereotypically feminine, masculine and neutral following Troles and

Schmid (2021). The results are gender-balanced datasets across 600 single sentences and 1100 contextual inputs, providing a measure of accuracy in translating gender.

Robustness to Character Noise This task evaluates how introducing word-level synthetic errors into source sentences affects the translation quality of an NMT model. We utilize the FLORES-200 devtest dataset (Costa-jussà et al., 2022), which allows us to evaluate the model's robustness to character perturbations across a wide range of directions. We implement three types of synthetic noise that have been previously used to stress NMT systems (Belinkov and Bisk, 2018; Peters and Martins, 2024):

- **swap:** For a selected word, two adjacent characters are swapped.
- **chardupe:** A character in the selected word is duplicated.
- **chardrop:** A character is deleted from the selected word.

A noise level parameter $\lambda \in [0, 1]$ controls the proportion of words in each sentence subjected to perturbations. Then, we evaluate the translation quality for each noise level using overlap and neural reference based metrics.

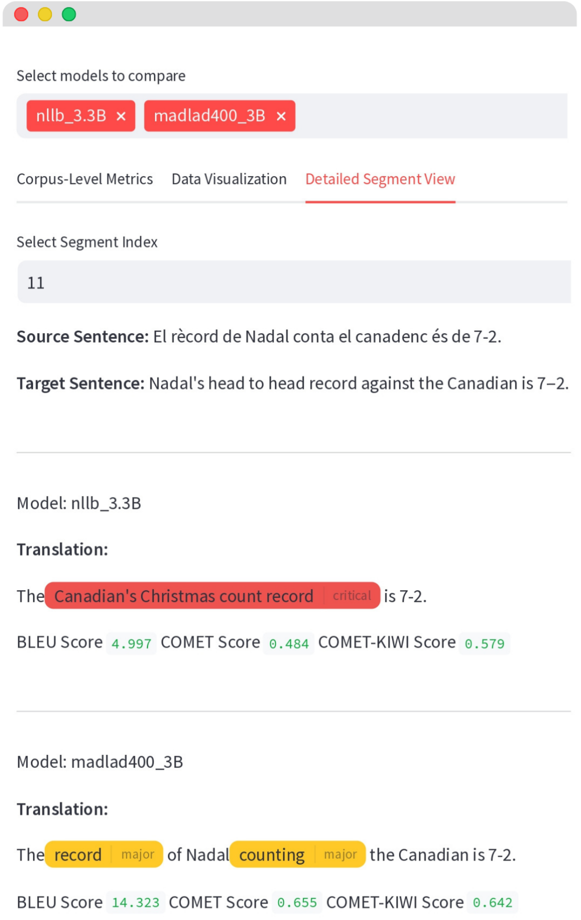
4 MT-LENS UI

The web user interface is organized into four main sections, each corresponding to a different MT task (Figure 2). It is built in Python using the Streamlit framework[3]. In this section, we describe the tools implemented in the user interface of MT-LENS and demonstrate its utility by evaluating two state-of-the-art NMT systems: madlad-400-3B (Kudugunta et al., 2024) and NLLB-3.3B (Costa-jussà et al., 2022), for the Catalan-to-English translation direction.

4.1 MT-LENS UI: Translation

4.1.1 Segment-by-Segment Comparison

MT-LENS allows users to analyze and compare translations across different systems. It first displays the source and target sentences for the selected segment, followed by the corresponding translations from the selected models (Figure 1).



Select Segment Index

11

Source Sentence: El rècord de Nadal conta el canadenc és de 7-2.

Target Sentence: Nadal's head to head record against the Canadian is 7–2.

Model: nllb_3.3B

Translation:

The Canadian's Christmas count record critical is 7-2.

BLEU Score 4.997 COMET Score 0.484 COMET-KIWI Score 0.579

Model: madlad400_3B

Translation:

The record major of Nadal counting major the Canadian is 7-2.

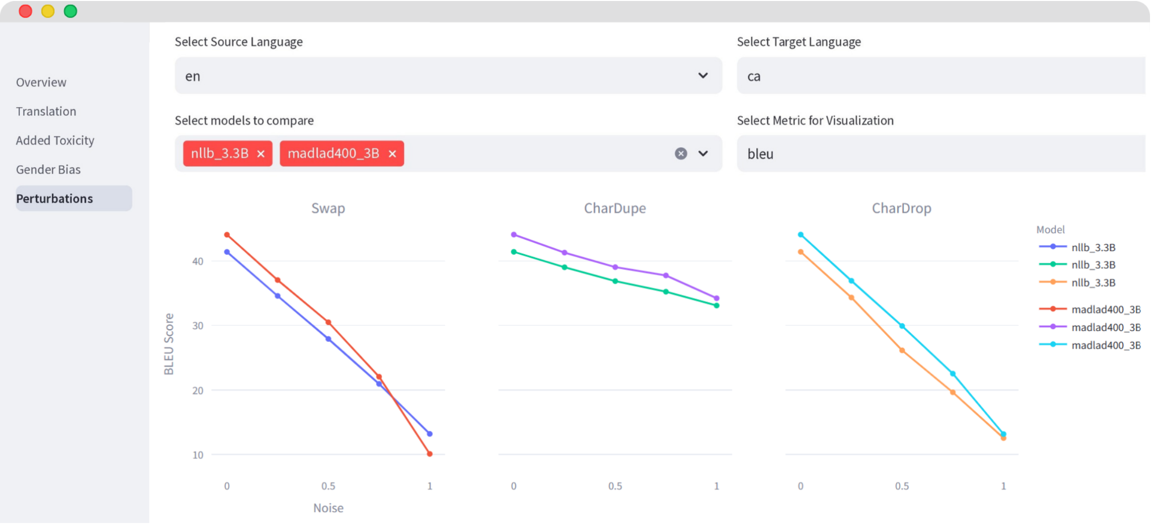
BLEU Score 14.323 COMET Score 0.655 COMET-KIWI Score 0.642

Figure 1: Segment comparison with error spans produced by madlad-400-3B and NLLB-3.3B systems.

Identifying and categorizing the errors made by NMT systems can be highly informative when comparing different models. In MT-LENS, if the XCOMET metric has been computed when evaluating a model, we use it to highlight error spans in a translation and marking them with different colors that indicate the severity of the error: red for critical errors, yellow for major errors, and blue for minor errors. We also provide individual segment scores for BLEU, COMET, and COMET-KIWI, which can be used to understand the translation's similarity to the reference text, its semantic similarity to the reference, and its semantic similarity to the source sentence respectively.

In Figure 1, we show an example of the segment-by-segment comparison page. We can see that the translation given by NLLB-3.3B has been categorized as critical by XCOMET while madlad-400-3B attains better results in individual metric scores, although it still produces two major errors.

[3] https://streamlit.io/



Select Source Language: en

Select Target Language: ca

Select models to compare: nllb_3.3B × madlad400_3B ×

Select Metric for Visualization: bleu

Figure 2: An image from the Perturbations page in the MT-LENS UI. Users can navigate between the following options: (1) Overview, (2) Translation, (3) Added Toxicity, (4) Gender Bias, and (5) Perturbations.

4.1.2 Segment-Length Analysis

Sentence length significantly influences system performance, with NMT systems often producing lower-quality translations for very long sentences (Koehn and Knowles, 2017). For analyzing the effect of sentence length on translation quality, MT-LENS offers interactive scatter plots, where the x-axis represents the number of words in a sentence, and the y-axis displays the corresponding score of that sentence for the selected metric.

4.1.3 Statistical Significance Testing

When comparing NMT systems, MT-LENS provides a visual interface for computing statistical significance testing through bootstrapped t-tests (Koehn, 2004) on BLEU, COMET and COMET-KIWI metrics. Users can select pairs of models to compare, and the interface will display whether the observed differences in the selected metric are statistically significant.

4.2 MT-LENS UI: **Added Toxicity**

Inspecting those segments that contain added toxicity can be particularly interesting when evaluating a NMT system using HOLISTICBIAS. Using MT-LENS UI, the user can see the obtained metrics aggregated at the system level and inspect the terms detected by ETOX for the selected model along a specific axis at the segment level.

4.3 MT-LENS UI: **Gender Bias**

Evaluating gender bias is crucial for developing fair and inclusive NMT systems. The MT-LENS UI offers a dedicated interface for assessing gender bias, showing aggregated metrics at the system level for the selected dataset. This interface is organized into two tabs, each corresponding to MUST-SHE and MMHB, respectively.

4.4 MT-LENS UI: **Perturbations**

Understanding how different translation systems handle input perturbations is crucial for assessing their robustness to real-world applications. MT-LENS UI, provides different visualizations to compare system performance for each type of noise evaluated. In Figure 2, we present an example of the Perturbations interface. The results show that madlad-400-3B exhibits greater robustness than NLLB-3.3B across all types of synthetic noise evaluated using the BLEU metric.

5 Conclusion

In this paper, we introduced MT-LENS, a framework designed to address existing gaps in MT evaluation by unifying various evaluation strategies. MT-LENS supports a diverse range of MT tasks, including traditional translation quality evaluation, gender bias detection, added toxicity, and robustness to character noise. By building upon the widely adopted LM-eval-harness library, MT-LENS provides seamless integration for evaluating both NMT and LLM-based models across various tasks. MT-LENS also offers a platform designed to provide insights into system performance. We believe MT-LENS has the potential to become the new adopted framework for evaluating NMT systems in the research community.

6 Limitations

Our tool is designed to provide a robust framework for researchers in machine translation to analyze various aspects of evaluation with ease. It is built to be adaptable and extendable, enabling the community to seamlessly incorporate new machine translation datasets. While standard evaluation metrics for machine translation are consistent across datasets, metrics for analyzing specific phenomena like gender bias and toxicity often require customization to suit the dataset. As a result, incorporating new metrics or datasets might occasionally require some additional effort or minor adjustments to the user interface. However, this is an area of active development, and we aim to implement methods in the future that will enhance flexibility and streamline these tasks even further.

7 Ethical Statement

Gender bias and toxicity in machine translation are multifaceted challenges that encompass a wide range of phenomena. The gender bias datasets used in this work primarily focus on evaluating coreference accuracy within a binary classification framework (male/female). For toxicity detection, we rely on the ETOX dataset, which identifies content deemed universally toxic, independent of context. While we acknowledge the limitations of these approaches, our objective is to represent widely recognized datasets for these tasks and contribute to a broader understanding of machine translation evaluation. This work does not aim to provide an exhaustive treatment of these complex issues but rather to offer a representative perspective.

8 Acknowledgements

This work has been promoted and financed by the Generalitat de Catalunya through the Aina Project.

This work has been supported by the Spanish project PID2021-123988OB-C33 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE.

This work is partially supported by DeepR3 (TED2021-130295B-C32) funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR.

This work was supported by EU Horizon 2020 project ELOQUENCE13[4] (grant number 101070558).

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the International Conference on Learning Representations*.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Laurent Bié, Aleix Cerdà-i Cucó, Hans Degroote, Amando Estela, Mercedes García-Martínez, Manuel Herranz, Alejandro Kohan, Maite Melero, Tony O'Dowd, Sinéad O'Gorman, Mārcis Pinnis, Roberts Rozis, Riccardo Superbo, and Artūrs Vasiļevskis. 2020. Neural translation for the European Union (NTEU) project. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 477–478, Lisboa, Portugal. European Association for Machine Translation.

[4] https://eloquenceai.eu/

Marta Costa-jussà, David Dale, Maha Elbayad, and Bokai Yu. 2024a. Added toxicity mitigation at inference time for multimodal and massively multilingual translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 360–372, Sheffield, UK. European Association for Machine Translation (EAMT).

Marta Costa-jussà, Mariano Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alexandre Mourachko, Christophe Ropers, and Carleigh Wood. 2024b. MuTox: Universal MUltilingual audio-based TOXicity dataset and zero-shot detector. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5725–5734, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Marta Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9570–9586, Singapore. Association for Computational Linguistics.

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. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Batya Friedman and Helen Nissenbaum. 1995. Minimizing bias in computer systems. In *Human Factors in Computing Systems, CHI '95 Conference Companion: Mosaic of Creativity, Denver, Colorado, USA, May 7-11, 1995*, page 444. ACM.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Javier García Gilabert, Carlos Escolano, and Marta Costa-jussà. 2024. ReSeTOX: Re-learning attention weights for toxicity mitigation in machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 37–58, Sheffield, UK. European Association for Machine Translation (EAMT).

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023a. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784, Toronto, Canada. Association for Computational Linguistics.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2023b. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Ondrej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. MT-ComparEval: Graphical evaluation interface for Machine Translation development. *The Prague Bulletin of Mathematical Linguistics*, 104:63–74.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Audrey Mash, Carlos Escolano, Aleix Sant, Maite Melero, and Francesca de Luca Fornaciari. 2024. Unmasking biases: Exploring gender bias in English-Catalan machine translation through tokenization analysis and novel dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17144–17153, Torino, Italia. ELRA and ICCL.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. Association for Computational Linguistics.

Ben Peters and André FT Martins. 2024. Did translation models get more robust without anyone even noticing? *arXiv preprint arXiv:2403.03923*.

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, Lisbon, Portugal. Association for Computational Linguistics.

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.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Craig Stewart, Luisa Coheur, and Alon Lavie. 2021. MT-Telescope: An interactive platform for contrastive evaluation of MT systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 73–80, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139, Bangkok, Thailand. Association for Computational Linguistics.

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024a. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA. Association for Computational Linguistics.

Beatrice Savoldi, Andrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024b. A prompt response to the demand for automatic gender-neutral translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 2: Short Papers, St. Julian's, Malta, March 17-22, 2024*, pages 256–267. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. “I’m sorry to hear that”: Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Craig Stewart, Ricardo Rei, Catarina Farinha, and Alon Lavie. 2020. COMET - deploying a new state-of-the-art MT evaluation metric in production. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 78–109, Virtual. Association for Machine Translation in the Americas.

Xiaoqing Ellen Tan, Prangthip Hansanti, Carleigh Wood, Bokai Yu, Christophe Ropers, and Marta R. Costa-jussà. 2024. Towards massive multilingual holistic bias. *Preprint*, arXiv:2407.00486.

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, Online. Association for Computational Linguistics.

Jonas-Dario Troles and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online. Association for Computational Linguistics.

Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine translation evaluation online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland. European Association for Machine Translation.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and André FT Martins. 2024. Watching the watchers: Exposing gender disparities in machine translation quality estimation. *arXiv preprint arXiv:2410.10995*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.