VERA: A Platform for Automatic and Human Evaluation of Machine Translation

Sofía García González[1 2], **Inés Quintana Raña**[1], **Jorge N. Afonso Cabido**[1], **Alberto Hernández Lado**[1], **German Rigau Claramunt**[2], **Sheila Castilho**[3]

[1]imaxin, Santiago de Compostela, Spain
{firstName.lastName}@imaxin.com
[2]University of the Basque Country (EHU), Donostia, Spain
german.rigau@ehu.eus
[3]SALIS/ADAPT Centre, Dublin City University, Dublin, Ireland
sheila.castilho@dcu.ie

Abstract

We present VERA, an easy-to-use platform for machine translation (MT) evaluation, combining both automatic metrics and the Multidimensional Quality Metrics (MQM) Core human evaluation framework in a single web environment. It supports reference-based metrics, multi-user annotation, corpus export, and PDF reports with automatic and human evaluation results, including their correlations.

1 Introduction

The rapid evolution of MT has led to an increasing need for robust and comprehensive evaluation platforms. While automatic metrics offer scalability and efficiency, human evaluation remains the gold standard for assessing MT quality. However, existing tools address these approaches separately. Notable examples include MATEO[1] (Vanroy et al., 2023) which focuses on automatic metrics via a web interface (BLEU, chrF, TER, COMET, BERTScore, and BLEURT); MT–LENS[2] (Gilabert et al., 2025) which extends to bias and robustness analysis beyond quality scores, and Pearmut[3] (Zouhar and Kocmi, 2026) which is specialised in human evaluation frameworks such as MQM, DA, and ESA. This fragmentation results in disconnected workflows. In this paper, we present VERA, a novel evaluation platform that seamlessly integrates automatic and human evaluation frameworks within a unified web environment. Uniquely, VERA further assesses the correlation between automatic and human evaluation results, enabling meta-evaluation and deeper analysis of MT evaluation methodologies. By bridging these perspectives, VERA provides more holistic and reliable evaluations of MT systems, supporting both research and industry needs.

VERA is being developed through an industrial PhD collaboration between imaxin, the University of the Basque Country (EHU), and the ADAPT Centre as a proprietary internal tool.[4]

2 Tool Description

VERA is a web-based platform composed of a Next.js[5] frontend and a FastAPI[6] backend. It provides multilingual support through next-intl[7] and integrates with Authentik[8] as an external identity provider, while NextAuth[9] manages session handling within the application. This architecture enables secure authentication and reliable access control for multiple concurrent users.

Automatic Evaluation MT outputs are evaluated using standard reference-based metrics—BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), TER (Snover et al., 2006) via SacreBLEU (Post, 2018)—, and COMET (Rei et al., 2020). Statistical significance between models is measured using bootstrap resampling (Koehn, 2004) for each metric.

[1]Source: https://mateo.ivdnt.org/
[2]Source: https://github.com/langtech-bsc/mt-evaluation
[3]Source: https://github.com/zouharvi/pearmut
[4]Although VERA is an internal tool, it is available via institutional licensing to universities and research groups for a fee. All the information is available at: https://imaxin.com/en/vera
[5]Source: https://github.com/vercel/next.js/
[6]Source: https://fastapi.tiangolo.com/
[7]Source: https://next-intl.dev/
[8]Source: https://goauthentik.io/
[9]Source: https://next-auth.js.org/

Human Evaluation VERA supports multiuser annotation following the MQM Core framework.[10] To make the human evaluation process more efficient, the platform integrates sampling strategies from Zouhar et al. (2025) to select informative evaluation subsets. Users can specify the proportion of the corpus to evaluate, after which the platform automatically selects the most representative segments.

Within the same interface, VERA also allows the creation and refinement of reference corpora. Existing references can be directly edited, and when no references are available, new ones can be generated either from scratch or by editing MT outputs.

Results Deployment Evaluation results (both automatic and human) are accessible directly within the platform, offering an interactive overview of MT model performance and metric correlations calculated using Pearson (Pearson, 1896) and Spearman (Spearman, 1904). A comprehensive PDF report can also be generated, and metric scores, reference datasets, and annotated corpora can be exported in CSV and JSON formats, enabling advanced analysis and seamless integration with external processing pipelines.

References

Gilabert, Javier García, Carlos Escolano, Audrey Mash, Xixian Liao, and Maite Melero. 2025. MT-LENS: An all-in-one Toolkit for Better Machine Translation Evaluation. In *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.

Koehn, Philipp. 2004. Statistical Significance Tests for Machine Translation Evaluation. In Lin, Dekang and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

[10]Source: `https://themqm.org/the-mqm-typology/`

Pearson, Karl. 1896. VII. Mathematical Contributions to the Theory of Evolution.—III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, (187):253–318.

Popović, Maja. 2017. chrF++: words helping character n-grams. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

Post, Matt. 2018. A Call for Clarity in Reporting BLEU Scores. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, et al., editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Snover, Matthew, 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, August 8-12. Association for Machine Translation in the Americas.

Spearman, Charles. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101.

Vanroy, Bram, 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.

Zouhar, Vilém and Tom Kocmi. 2026. Pearmut: Human Evaluation of Translation Made Trivial. *arXiv preprint arXiv:2601.02933*.

Zouhar, Vilém, Peng Cui, and Mrinmaya Sachan. 2025. How to Select Datapoints for Efficient Human Evaluation of NLG Models? *Transactions of the Association for Computational Linguistics*, 13:1789–1811.