The Journal of Open Source Software

Jury: A Comprehensive Evaluation Toolkit

Devrim Cavusoglu[1,2*¶], **Secil Sen**[1,3*], **Ulas Sert**[1], **and Sinan Altinuc**[1,2]

1 OBSS AI **2** Middle East Technical University **3** Bogazici University ¶ Corresponding author * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Chris Vernon

Reviewers:

- @evamaxfield
- @KennethEnevoldsen

Submitted: 23 January 2024
Published: unpublished

Summary

Evaluation plays a critical role in deep learning as a fundamental block of any prediction-based system. However, the vast number of Natural Language Processing (NLP) tasks and the development of various metrics have led to challenges in evaluating different systems with different metrics. To address these challenges, we introduce jury, a toolkit that provides a unified evaluation framework with standardized structures for performing evaluation across different tasks and metrics. The objective of jury is to standardize and improve metric evaluation for all systems and aid the community in overcoming the challenges in evaluation. Since its open-source release, jury has reached a wide audience and is publicly available.

Statement of need

NLP tasks possess inherent complexity, requiring a comprehensive evaluation of model performance beyond a single metric comparison. Established benchmarks such as WMT (Barrault et al., 2020) and GLUE (Wang et al., 2018) rely on multiple metrics to evaluate models on standardized datasets. This practice promotes fair comparisons across different models and pushes advancements in the field. Embracing multiple metric evaluations provides valuable insights into a model's generalization capabilities. By considering diverse metrics, such as accuracy, F1 score, BLEU, and ROUGE, researchers gain a holistic understanding of a model's response to never-seen inputs and its ability to generalize effectively. Furthermore, task-specific NLP metrics enable the assessment of additional dimensions, such as readability, fluency, and coherence. The comprehensive evaluation facilitated by multiple metric analysis allows for trade-off studies and aids in assessing generalization for task-independent models. Given these numerous advantages, NLP specialists lean towards employing multiple metric evaluations.

Although employing multiple metric evaluation is common, there is a challenge in practical use because widely-used metric libraries lack support for combined and/or concurrent metric computations. Consequently, researchers face the burden of evaluating their models per metric, a process exacerbated by the scale and complexity of recent models and limited hardware capabilities. This bottleneck impedes the efficient assessment of NLP models and highlights the need for enhanced tooling in the metric computation for convenient evaluation. In order for concurrency to be beneficial at a maximum level, the system may require hardware accordingly. Having said that, the availability of the hardware comes into question.

The extent of achievable concurrency in NLP research has traditionally relied upon the availability of hardware resources accessible to researchers. However, significant advancements have occurred in recent years, resulting in a notable reduction in the cost of high-end hardware, including multi-core CPUs and GPUs. This progress has transformed high-performance computing resources, which were once prohibitively expensive and predominantly confined to specific institutions or research labs, into more accessible and affordable assets. For instance, in BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019), it is stated that they leveraged the

training process by using powerful yet cost-effective hardware resources. Those advancements show that the previously constraining factor for hardware accessibility has been mitigated, allowing researchers to overcome the limitations associated with achieving concurrent processing capabilities in NLP research.

To ease the use of automatic metrics in NLG research, several hands-on libraries have been developed such as *nlg-eval* (Sharma et al., 2017) and *datasets/metrics* (Lhoest et al., 2021) (now as *evaluate*). Although those libraries cover widely-used NLG metrics, they don't allow using multiple metrics in one go (i.e. combined evaluation), or they provide a crude way of doing so if they do. Those libraries restrict their users to compute each metric sequentially if users want to evaluate their models with multiple metrics which is time-consuming. Aside from this, there are a few problems in the libraries that support combined evaluation such as individual metric construction and passing compute time arguments (e.g. n-gram for BLEU (Papineni et al., 2002)), etc. Our system provides an effective computation framework and overcomes the aforementioned challenges.

We designed a system that enables the creation of user-defined metrics with a unified structure and the usage of multiple metrics in the evaluation process. Our library also exploits *datasets* package to promote the open-source contribution; when users implement metrics, the implementation can be contributed to the *datasets* package. Any new metric released by *datasets* package will be readily available in our library as well.

Acknowledgements

We would also like to express our appreciation to Cemil Cengiz for fruitful discussions.

References

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., … Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). *Proceedings of the Fifth Conference on Machine Translation*, 1–55. https://aclanthology.org/2020.wmt-1.1

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., Platen, P. von, Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., … Wolf, T. (2021). Datasets: A community library for natural language processing. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 175–184. https://doi.org/10.18653/v1/2021.emnlp-demo.21

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Sharma, S., El Asri, L., Schulz, H., & Zumer, J. (2017). Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, *abs/1706.09799*. http://arxiv.org/abs/1706.09799

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the*

2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 353–355. https://doi.org/10.18653/v1/W18-5446

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf