HERMeS:
Human Evaluation & Ranking of Multiple Systems

Rex VanHorn
Institute for Artificial Intelligence,
University of Georgia, Athens, Georgia, USA
ORCID ID: 0000-0002-3792-427
rex.vanhorn@uga.edu

Abstract

HERMeS is a lightweight human evaluation platform designed to streamline human evaluation and comparison of multiple MT systems across large translation sets. As a complement to existing evaluation tools, HERMeS focuses specifically on scalable comparison of many anonymized systems through a hybrid ranking and direct assessment workflow, using a practical approach that reduces cognitive load while maintaining data quality, security, and integrity.

1 Introduction

Reliable evaluation remains one of the central challenges in machine translation research. While automatic metrics provide convenient approximations of translation quality, human evaluation remains the gold standard. However, prior work has highlighted ongoing challenges in human MT evaluation, including annotator effort, consistency, and the difficulty of defining translation quality (Castilho, 2021). On the other hand, modern research settings frequently require comparisons across many systems, including commercial translation engines and large language models. In such scenarios, evaluation tools must balance methodological rigor with practical usability for human annotators.

This paper presents a human evaluation platform, HERMeS, designed to support scalable and reproducible, sentence-level evaluation of multiple MT systems, combining two complementary evaluation approaches: categorical ranking through bucket assignments and fine-grained direct-assessment scoring, thereby reducing evaluator fatigue while keeping evaluation quality consistent.

2 System Design

HERMeS follows a two-step workflow consisting of bucket assignment (coarse ranking) and direct assessment (numeric scoring). It was developed in support of our ongoing study and seeks to reduce the overhead of human evaluation while maintaining quality and integrity. The system presents source sentences alongside MT-/LLM-generated translations to efficiently collect human judgments.

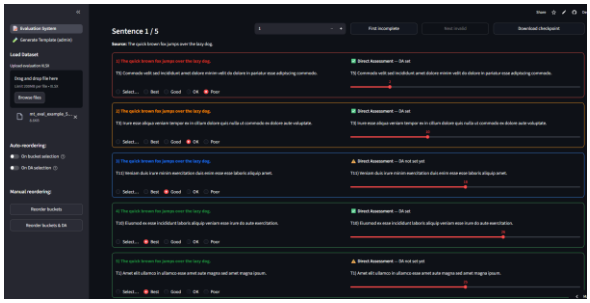

Figure 1: Screenshot of HERMeS

2.1 Evaluation Interface

The evaluators load the evaluation spreadsheet into HERMeS' user interface, through which one source sentence is displayed at a time as a translation card. The translation card shows the source sentence, translation, an evaluation bucket on the left, and a direct-assessment slider on the right. Each translation card allows the evaluator to provide two types of judgments (required).

Bucket Assignment

Evaluators first assign translations to one of four relative quality categories: Best, Good, OK, or Poor (these names are configurable). Bucketed ranking provides a low-effort entry point for evaluation by allowing annotators to make coarse

comparative judgments before assigning precise scores. This reduces early-stage decision complexity and helps establish a stable internal ranking prior to fine-grained assessment. HERMeS can automatically group all translations by bucket upon ranking, or upon completion, such that all translations assigned to a common quality category are displayed together.

Direct Assessment (DA)

Evaluators then provide a numerical quality score on a fine-grained scale of the research team's choice; the range is configurable. In our current research, each bucket corresponds to a fixed scoring range of size 7 (e.g., 1–7 for Poor, 8–14 for OK, etc.). This score represents the perceived adequacy and fluency of the translation relative to the source sentence. The system can order the DA scores within a bucket upon assessment, and upon the evaluator's button click. Additionally, evaluators can make notes about the source sentence, its translations, or the evaluation itself, per source sentence.

While adequacy and fluency are distinct dimensions, prior work has shown they are often correlated in practice. In lieu of separate dimension scoring, HERMeS offers a hybrid workflow that integrates bucketed ranking and direct assessment, capturing both relative and absolute quality while reducing annotator burden, in a format that easily lends itself to final review. Note that scores can be reassigned at any time, and the system is not constrained by dataset size or text length.

2.2 Data Integrity

Source data are randomized, anonymized, and integrity-checked prior to evaluation. Using hash-based methods to validate the inputs and outputs enables the detection of tampering and the reliable merging of distributed annotations.

3 Evaluation Workflow

HERMeS' annotation workflow is designed to minimize evaluator friction while ensuring coherent data collection. Evaluators proceed through the evaluation sentences sequentially, completing rankings for each source sentence's translations. The interface supports automatic reordering of translations as bucket assignments are made, or upon completion, which helps maintain coherent rankings. Evaluators then provide direct assessment scores for translations in each of the buckets, with automatic or on-demand ordering, if desired. Evaluators then have the option to review all assessments together.

HERMeS is a stateless application that requires no login, and is implemented as a web-based interface built using the Streamlit framework,[2] and runs on the Streamlit Community Cloud platform.[3] Accordingly, the results are not saved by the system, but rather evaluators are reminded and encouraged to save their work by downloading 'checkpoints', which can be exported and saved at any time in spreadsheet format, thereby allowing annotation sessions to be paused and resumed without data loss, and with no cost to the researchers or evaluators. The software is open-source and available on GitHub.[4]

Upon completion of their annotations, evaluators save the final checkpoint spreadsheet and return it to the research team, which can then de-anonymize and de-randomize the results for analysis.

Appraise (Federmann, 2012) established a strong general-purpose foundation for manual MT assessment. HERMeS is designed as a complement to fine-grained annotation systems like Appraise, targeting large-scale, direct-assessment comparison of many anonymized systems in a single workflow. It was designed to reduce evaluator fatigue, provide distributed, spreadsheet-based portability, and offer built-in, hash-based integrity checks, without the need for coding or technical expertise. It was not designed to capture fine-grained error categories and is not intended for detailed error analysis. In an evaluation study using HERMeS (13 systems, 100 sentences; 1,300 translations), preliminary qualitative feedback suggests that a bucket-first workflow may reduce perceived effort and improve alignment and consistency.

4 References

Castilho, S. (2021, April). Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In Proceedings of the workshop on human evaluation of NLP systems (HumEval) (pp. 34–45).

Federmann, C. (2012). Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output. Prague Bull. Math. Linguistics, 98, 25–36.

[2] https://streamlit.io

[3] https://mt-eval.streamlit.app/?type=admin

[4] https://github.com/RexVH/St-MT-Evaluation-app