# Do Automatic Metrics Capture What Humans See in Machine Translation Evaluation?

Authors: Venna Mounika, University of Hyderabad, India
C.Rajyarama, University of Hyderabad, India
Abstract ID: 457
Event: Linguistics 2026 Paris
Subject: 19. Computational linguistics
Presenter Name: Venna Mounika

## Abstract

Machine Translation (MT) systems and Large Language Models (LLMs) now produce fluent texts and improve rapidly. However, this rapid growth creates an important question: Can Automatic evaluation tools be trusted or do we still need Human judges? This study examines whether Automatic evaluation methods can accurately find translation errors in English-Telugu translation. Telugu is an agglutinative language in which words are formed by joining multiple suffixes to the root, marking case, number, person, gender, tense, and modality, and in which a modifying clause comes before the noun it describes. One wrong suffix can change who does what to whom or turn "must" into "need not" (a shift in deontic modality). This raises two important questions: (1) Does smooth and natural-looking machine output hide serious grammatical and meaning errors? (2) Can Automatic evaluation methods find complex word-structure errors without help from Human judges? We hypothesise that single-score metrics miss them and that error-based metrics need human judgments to set how serious each error type is in each domain.

In the study, two models are compared: IndicTrans2 (designed for Indian languages) and GPT-4o (a decoder-only LLM). A total of 19,000 English sentences is considered for translation from four formal domains: Agriculture, Health care, Governance, and Sports. A two-tier quality assessment is performed on the translated output. The Automatic evaluation tier used COMET-QE, a neural model that provides a single overall score and DELTA, an LLM-based tool adapted from GEMBA-MQM that labels each error by type and severity. DELTA sorts errors into six dimensions: Accuracy, Lexical, Morphological, Sentence structure, Fluency and Style and Orthography. In the Human evaluation tier, expert native Telugu annotators evaluated a stratified sample, with separate sets for fitting domain-specific weights and for validation. Agreement between each metric and human judgments was measured using rank and linear correlations, with significance testing.

Our findings show that natural-looking translations often hide major errors. GPT-4o writes fluent sentences but it often drops or mixes up Telugu case markers and verb endings and broken pre-nominal participial clauses. In Healthcare and Governance texts, it sometimes reverses crucial rules or adds false details. IndicTrans2 handles word endings better because its tokenizer was built specifically for Indian Languages. The Automatic metrics COMET-QE provides general scores but fails to flag severe grammatical errors. In contrast, DELTA successfully identified the errors and the weights fitted to Human judgments agreed more closely with the annotators than domain-agnostic scoring did. The annotators agreed substantially on the severity of structural errors.

The results show that automatic metrics alone are not enough for typologically distinct language pairs like English and Telugu. When Automatic metrics give only a single score, they miss hidden grammatical mistakes that can cause serious real-world problems in domains like healthcare. LLM-based evaluators make error analysis scalable but they rely on Human judgments to decide how much each error matters in each domain. We therefore suggest using Human experts for calibration and audit but not for every new system.