Translation Quality Assessment: A Concise Review of Methods,Metrics, and Developments[1]

Res. Hussein Ali Al-Mahmood

Asst. Prof. Abdulsalam Al-Ogaili

Department of Translation / College of Arts / University of Basrah

Abstract

Translation quality assessment has evolved significantly from early human-based assessment methods to sophisticated neural metrics that leverage deep learning architectures. This concise review examines the development, methodologies, and performance of evaluation approaches across four decades of machine translation research, and surveys three primary evaluation paradigms: human evaluation methods including Direct Assessment, Multidimensional Quality Metrics, and Error Span Annotation; widely used automatic lexical metrics such as BLEU, METEOR, and TER; and state-of-the-art neural evaluation metrics including COMET, BLEURT, and GEMBA-MQM. The paper also explores meta-evaluation techniques, highlighting the transition from Pearson, Spearman and Kendall's correlations to more robust measures like Soft Pairwise Accuracy. Following a systematic examination of evaluation methodologies, experimental findings, and comparative analyses, we apply these frameworks to Al-Mahmood's English⇔Arabic evaluation dataset to validate the key strengths and limitations of current approaches.

Keywords: machine translation evaluation, automatic metrics, translation quality assessment, human annotation.

Received: 06/16/2025 **Accepted: 15/07/2025**

تقييم الترجمة الآلية: مراجعة مقتضبة للطرائق والمقاييس والتطورات

الباحث حسين علي المحمود
الأستاذ المساعد الدكتور عبد السلام عبد المجيد العكيلي
قسم الترجمة / كلية الآداب / جامعة البصرة

المستخلص

تطور مجال تقييم الترجمة الآلية بشكل كبير من طرق التقييم البشرية المبكرة إلى المقاييس العصبية المتطورة التي تستفيد من بنى التعلم العميق. تستعرض هذه المراجعة الموجزة تطور منهجيات التقييم وأدائها على مدى أربعة عقود من أبحاث الترجمة الآلية وتستعرض ثلاثة أنماط تقييم أساسية: طرق التقييم البشري بما في ذلك التقييم المباشر (DA) ومقاييس الجودة متعددة الأبعاد (MQM) وتوضيح نطاق الأخطاء (ESA)؛ والمقاييس المعجمية التلقائية واسعة الاستخدام مثل BLEU و METEOR ومعدل تحرير الترجمة (TER)؛ وأحدث مقاييس التقييم العصبية بما في ذلك COMET و BLEURT و GEMBA-MQM. تستكشف المقالة أيضًا تقنيات التقييم التلوي، وتسلط الضوء على الانتقال من ارتباطات بيرسون وسبيرمان وكيندال إلى مقاييس أكثر فاعلية مثل الدقة الزوجية الناعمة (SPA). بعد إجراء مراجعة منهجية لطرائق التقييم والنتائج التجريبية والتحليلات المقارنة، نطبق هذه الأطر على مجموعة بيانات التقييم الإنجليزية العربية التي نفذها المحمود (٢٠٢٥) للتحقق من نقاط القوة والقيود الرئيسية للنهج الحالية.

كلمات مفتاحية: التعليق البشري، تقييم الترجمة الآلية، المقاييس التلقائية، تقييم جودة الترجمة.

تاريخ الاستلام: ٢٠٢٥/٠٦/١٦ **تاريخ القبول: ٢٠٢٥/٠٧/١٥**

1.Introduction

The evaluation of machine translation (MT) quality has been a fundamental challenge since the inception of automated translation systems. MT evaluation (also known as translation quality assessment) serves as both a diagnostic tool for system improvement and a comparative benchmark for research advancement. The field has witnessed a remarkable transformation from early human-based evaluation protocols to sophisticated neural metrics that leverage deep learning architectures to predict translation quality. The importance of robust MT evaluation methodologies became evident by the infamous ALPAC report (National Research Council, 1966), which highlighted the critical need for systematic evaluation approaches in MT research. This report underscored the complexity of evaluation and led to increased focus on developing reliable evaluation methodologies that could support the advancement of MT systems.

Early MT evaluation approaches primarily relied on human evaluators who assessed translations according to measures such as adequacy and fluency (Castilho et al., 2018). Adequacy measures the extent to which the translation conveys the meaning of the source text into the target language, focusing on correctness and completeness, while fluency assesses how well the translation adheres to the rules and norms of the target language, emphasizing grammatical correctness, naturalness, and readability. These foundational concepts continue to influence modern evaluation frameworks, even as the field has embraced increasingly automated approaches.

Although novel error analysis methods continue to emerge, such as the eclectic model proposed by Musa and Al-Maryani (2021) based on Riccardi's (1999) and Naj'a and Abu-Mighnim's (2012) models, in addition to classical works such as Juliane House's model (1977), these models are not specifically designed for evaluating MT output. This limitation has been effectively addressed by dedicated metrics and frameworks that have been specifically developed and designed to assess MT performance.

Modern evaluation frameworks such as MQM, while more comprehensive than adequacy and fluency measurement, also face many challenges. One of the main shortcomings of human evaluation methodologies, including MQM, is that evaluation performed by human

annotators is slow, costly, and particularly time-consuming. This makes it difficult to assess translation quality at scale. As a result, there is a growing interest in automated methods of evaluation, such as Quality Estimation (QE), which can provide a more efficient and scalable way to assess translation quality (Silva et al., 2024). Furthermore, it is crucial to emphasize the influence of human subjectivity and potential errors in judgment. For instance, several works (see Song et al., 2025) have shown that inter-annotator agreement on high-quality WMT MQM datasets yields low to modest correlation.

The introduction of statistical MT systems further necessitated automatic evaluation metrics that could provide rapid feedback during system development. This led to the development of string-matching metrics such as BLEU (Papineni et al., 2002), which revolutionized the field by enabling large-scale comparative evaluations. However, the limitations of purely lexical approaches became apparent as translation systems grew more sophisticated, prompting the development of neural evaluation metrics that leverage semantic understanding.

Contemporary MT evaluation research faces multiple challenges such as "benchmark coverage, particularly regarding dialectal variations, code-switching phenomena, and specialized domains where human expertise is required for accurate assessment" (H. Li et al., 2025, p. 4). The rise of large language models (LLMs) and neural MT has both created new opportunities for evaluation. The integration of these models in MT evaluation have primarily benefited high-resource languages. They have demonstrated impressive performance by capturing deeper semantic relationships in translations, but their effectiveness diminishes for underrepresented languages. This is due to the scarcity of high-quality training and evaluation data (S. Li et al., 2025, p. 1).

This review examines the development of MT evaluation methodologies from their origins to current state-of-the-art approaches. We review human evaluation methods, automatic lexical metrics, neural evaluation approaches, meta-evaluation techniques, and conclude our review with an analysis of the key trends in English⇔Arabic evaluation by inspecting comparative performance of the different types of metrics in Al-Mahmood's (2025) latest dataset. Through all this, we aim to provide translators with a simple understanding of the complicated field of MT evaluation.

2. Methods and Types of Machine Translation Evaluation

The systematic evaluation of MT quality requires well-defined methodologies that can accommodate different use cases and evaluation objectives. White (2003) identified three fundamental evaluation methods that continue to structure MT evaluation approaches today. These methods differ in the information available to evaluators and the types of judgments they enable. These methods are: output only, input and output, and input and output with reference.

The first method involves judging target-language output without access to the source text. This approach focuses purely on the fluency and naturalness of the translation in the target language, making it suitable for assessing whether translations meet target language quality standards. However, this method cannot evaluate adequacy or meaning preservation, limiting its applicability in comprehensive quality assessment. The second method provides evaluators with both source and target texts, enabling assessment of meaning preservation and adequacy. This approach allows for more comprehensive evaluation but requires bilingual evaluators and may be influenced by the evaluator's proficiency in both languages. The third method includes a high-quality human reference translation alongside the source and MT output, which enables more structured comparison and can improve evaluation consistency by providing a quality benchmark. However, it introduces potential bias toward the reference translation style and may not account for legitimate translation variations (White, 2003).

White (2003) further categorized evaluation approaches into six distinct types, each serving different objectives in the development and deployment of MT systems:

1- Feasibility evaluation addresses fundamental questions about system viability in early research stages. This type of evaluation determines whether a proposed approach can achieve basic translation functionality and is typically employed during initial system development phases.

2- Internal evaluation examines specific system components through both glass-box (white-box) and black-box testing methodologies (Quah, 2006). Glass-box evaluation provides access to internal system states and intermediate representations, enabling detailed analysis of component performance. Black-box evaluation treats the system as an opaque entity,

focusing on input-output relationships without examining internal mechanisms.

3- Declarative evaluation provides objective performance judgment against predetermined criteria and benchmarks. This approach enables systematic comparison across systems and tracking of progress over time through standardized metrics and test sets.

4-Usability evaluation focuses on user experience and interface quality, examining how effectively end users can interact with translation systems. This type considers factors such as user satisfaction, task completion rates, and interface design effectiveness.

5- Operational evaluation assesses cost-effectiveness and practical considerations for organizational deployment. This includes analysis of computational requirements, maintenance costs, integration complexity, and return on investment for potential users.

6- Comparison evaluation enables side-by-side system analysis while controlling for specific error types and evaluation conditions. This approach is essential for research advancement and system selection in practical applications.

3. Human Evaluation of Machine Translation

Human evaluation remains the gold standard for MT evaluation, providing nuanced judgments that capture semantic, pragmatic, and stylistic aspects of translation quality. The evolution of human evaluation methodologies reflects ongoing efforts to improve reliability, consistency, and scalability while maintaining the depth of analysis that only human evaluators can provide. Early human evaluation approaches established foundational concepts that continue to influence contemporary MT evaluation. The ALPAC study (National Research Council, 1966) employed 36 Russian sentences with both human and MT systems, utilizing monolingual and bilingual raters to assess translation quality in English. This landmark study found a high correlation between intelligibility and fidelity measures, establishing the importance of these dimensions in quality assessment.

DARPA Machine Translation Evaluation Program (1992-1994) represented a significant advancement in systematic human evaluation methodology. The methodology of this program decomposes subjective human assessments into a large sample of small units, focusing on

separate evaluations for adequacy, informativeness, and fluency on a five to one degrading scale, where five indicates that all of the meaning is present, and one indicates that little or none of the meaning is present. This approach established the practice of using separate scales for different quality dimensions and provided a template for subsequent evaluation campaigns (White & O′Connell, 1996). *EAGLES evaluation* framework (Sparck Jones & Galliers, 1995) further refined scoring evaluation by providing structured guidelines for assessment procedures and quality criteria. These early efforts established best practices for human evaluation design, including the importance of clear evaluation guidelines, adequate evaluator training, and systematic quality control procedures.

Recognition of the limitations inherent in absolute scoring led to the development of ranking-based evaluation methods. Koehn and Monz (2006) pioneered manual evaluation approach involving 200-300 sentences per system, comparing 5 randomly selected systems through ranking procedures. The evaluation involved presenting translations to the judges in the form of ranking, which demonstrated improved consistency compared to absolute scoring methods. Binary system comparison ranking (Vilar et al., 2007) also emerged as a particularly effective approach, enabling evaluators to make more reliable judgments by focusing on pairwise comparisons rather than absolute quality assessments. This methodology reduces cognitive load on evaluators while improving inter-annotator agreement and evaluation reliability. The research by Vilar et al. (2007) and Duh (2008) also provided empirical evidence that ranking achieves higher correlation with human judgments than traditional scoring approaches. This finding led to the adoption of ranking methodologies in major evaluation campaigns, including the *Workshop on Machine Translation* (WMT), which has employed ranking as its primary evaluation approach since 2008 (Freitag et al., 2021).

The limitations of ranking approaches for large-scale evaluations prompted the development of *Direct Assessment* (DA) methodologies. Graham et al. (2013) introduced DA with continuous scales, comparing five-point scales versus 1-100 point scales. Their research demonstrated that continuous scale allows scores to be standardized to eliminate individual judge preferences, resulting in higher levels of inter-annotator consistency. DA has been the official evaluation methodology for WMT since 2017 (Freitag et al., 2021), reflecting its effectiveness in large-scale

evaluation scenarios. The approach relies on crowd-sourced annotators who provide absolute quality judgments on continuous scales, enabling efficient evaluation of multiple systems across various language pairs.

However, DA faces significant limitations that researchers continue to address. As noted by Birch et al. (2016), DA relies on a large number of crowd-sourced annotators and its lack of granularity provides a single, opaque score that may not capture the complexity of translation quality. The method requires extensive quality control procedures to ensure reliable results and may not provide sufficient detail for diagnostic purposes. Alternative approaches such as *Human UCCA-based MT Evaluation* (HUME) (Birch et al., 2016) offer more granular error analysis using bilingual annotators, but such methods are less widely adopted due to increased complexity and resource requirements. The trade-off between scalability and granularity remains a central challenge in human evaluation design.

Recognition of the multifaceted nature of translation quality led to the development of *Multidimensional Quality Metrics* (MQM) by Lommel et al. (2013). MQM provides a comprehensive error typology that enables detailed analysis of translation problems across multiple dimensions including accuracy, fluency, terminology, style, and locale-specific issues. The MQM framework employs a hierarchical error classification system with associated severity weights, enabling both diagnostic analysis and overall quality scoring. This approach provides valuable insights for system improvement while maintaining compatibility with comparative evaluation needs. The framework's flexibility allows adaptation to specific domains and use cases while maintaining systematic evaluation procedures. Segment-level adaptations of MQM, such as SQM (Freitag et al., 2021), have also demonstrated effectiveness in shared task evaluations. These approaches balance the need for detailed error analysis with practical constraints on evaluation time and resources, making them suitable for both research and industrial applications.

Error Span Annotation (ESA), introduced by Kocmi, Zouhar, et al. (2024), represents a recent advancement in human evaluation methodologies. ESA requires annotators to highlight specific error spans in translations and assign severity ratings, providing more precise localization of translation problems compared to sentence-level assessment approaches. This methodology offers greater flexibility than MQM while maintaining focus on specific error identification. ESA

enables detailed analysis of error patterns and distributions while providing data suitable for training automatic evaluation metrics. The approach addresses limitations of both coarse-grained sentence-level evaluation and complex multi-dimensional frameworks. While it does have several limitations (e.g., simple error types, potential scoring variability due to its subjective nature, etc.), Kocmi, Zouhar, et al. (2024) assert that it is more cost-effective than MQM for annotation, and produces annotations that are more closely aligned with human judgment compared to DA+SQM, due to it being less susceptible to the impact of fluency variations. Consequently, ESA has been adopted in the WMT24 in tasks like the General Machine Translation Shared Task (Kocmi, Avramidis, et al., 2024), Metrics Shared Task (Freitag et al., 2024), among others.

4. String Matching-based Automatic Metrics

The development of automatic evaluation metrics represented a paradigm shift in MT evaluation, enabling rapid evaluation of large numbers of translations and supporting iterative system development. Initial automatic evaluation approaches employed basic precision, recall, and F-measure calculations adapted from information retrieval methodologies (Koehn, 2010). These metrics provided objective, reproducible assessments but suffered from significant limitations in capturing translation quality nuances. As noted by Daelemans & Hoste (2009), these approaches were unable to provide meaningful fine-grained analysis of translation quality.

One of the early lexical metrics is BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), which marked a significant advancement in automatic evaluation methodology. BLEU employs n-gram matching between machine translations and reference translations, incorporating a brevity penalty to address length differences. The metric ranges from 0 to 1, with higher scores indicating greater similarity to reference translations. BLEU's effectiveness stems from its use of multiple n-gram orders (typically 1-4 grams) combined with geometric averaging, which provides sensitivity to both lexical choice and local word order. The brevity penalty prevents systems from achieving high scores through excessively short translations, addressing a significant limitation of pure precision-based approaches.

Extensive evaluation in NIST and WMT campaigns demonstrated BLEU's utility for system comparison and development. As documented

by Daelemans and Hoste (2009), BLEU scores can effectively highlight frequent mistranslations and omissions while showing high correlation between human judgments of informativeness and the normalized variation (N) score. However, BLEU faces significant limitations that have prompted ongoing research into alternative metrics. Its reliance on exact lexical matching fails to recognize legitimate translation variations and synonyms. Additionally, it shows inconsistent performance across different translation paradigms and may not adequately reflect improvements in neural machine translation systems (Koehn, 2020).

METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee & Lavie, 2005) addressed several limitations of BLEU by incorporating recall alongside precision and adding support for stemming, synonymy, and paraphrase matching. This approach provided more flexible matching criteria while maintaining focus on alignment between machine and reference translations. The metric's emphasis on recall helped address BLEU's bias toward shorter translations while the semantic matching capabilities improved recognition of translation variations. Its alignment-based approach also provided better handling of word order differences compared to purely n-gram-based metrics. However, despite these improvements, METEOR faced significant computational complexity challenges (Koehn, 2010) and limitations in combining multiple reference translations (Olive et al., 2011). The metric's reliance on external resources for synonymy and paraphrase detection also limited its applicability across diverse language pairs and domains.

Translation Edit Rate (TER) (Snover et al., 2006) introduced an edit-distance-based approach to automatic evaluation, measuring the minimum number of edits normalized by reference length required to transform a machine translation into a reference translation. TER considers four edit operations: insertions, deletions, substitutions, and phrase shifts. This focus on edit operations provided intuitive interpretation and demonstrated good correlation with human judgments in various evaluation campaigns. TER's explicit modeling of phrase movements addressed word order differences more effectively than n-gram-based approaches, making it particularly suitable for evaluating translations between languages with different syntactic structures.

TERp (Snover et al., 2009) extended TER by incorporating synonymy and stemming in alignment calculations, improving robustness to lexical variations. These enhancements addressed key limitations of the original

TER metric while maintaining its computational efficiency and interpretability. However, both TER and TERp remained limited by their reliance on single reference translations and exact matching requirements. The metrics also struggled with translations that employed significant restructuring while preserving meaning, often penalizing valid translation alternatives.

As noted earlier, one of the primary criticisms of lexical MT evaluation metrics is their limited scope to the lexical dimension. These metrics fail to encompass deeper linguistic information, such as syntactic and semantic aspects. This can make them less effective in evaluating MT systems that use different paradigms or have different lexicons (Giménez & Màrquez, 2007). However, despite the limitations of traditional lexical-based metric, they are considerably more sensitive to certain types of errors that neural metrics often struggle with (see Glushkova et al., 2023).

5. Neural Evaluation Metrics

The advent of deep learning and neural machine translation systems necessitated corresponding advances in evaluation methodologies. Neural evaluation metrics leverage learned representations and semantic understanding to provide more sophisticated assessment of translation quality, moving beyond the surface-level string matching that characterized earlier automatic metrics.

As noted by Lee et al. (2023), the field experienced initial stagnation followed by rapid growth in neural metric development. These metrics differ fundamentally from lexical approaches by incorporating semantic understanding, contextual awareness, and learned quality judgments. However, they also introduce new challenges related to computational requirements, domain adaptation, and interpretability.

BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) (Sellam et al., 2020) represented a significant advancement in neural evaluation metrics by leveraging BERT's (Devlin et al., 2018) pretrained representations adapted for translation evaluation. The metric employs a multi-stage training approach involving pretraining on synthetic data followed by fine-tuning on human judgment data. It demonstrates consistently higher correlation with human judgments across various datasets and tasks compared to traditional lexical metrics. The metric's success stems from its ability to

capture semantic similarity and contextual appropriateness through transformer-based representations, enabling more nuanced quality assessment. However, BLEURT faces robustness challenges, particularly regarding adversarial examples and domain adaptation. Research revealed robustness defects and susceptibility to adversarial translations in BLEURT, highlighting the need for more robust training procedures and evaluation protocols (Yan et al., 2023).

COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020) introduced multiple model variants designed for different evaluation scenarios: DARR Ranker for system ranking, MQM Estimator for error detection, and HTER Estimator for edit distance prediction. This multi-model approach enables targeted evaluation for specific use cases while maintaining state-of-the-art performance. Experimental results demonstrated that DARR Ranker model outperforms the two Estimators in seven out of eight language pairs and outperforms BLEU and CHRF, as well as YISI-1 and BERTSCORE (Rei et al., 2020). The metric's superior performance across diverse language pairs established it as a leading neural evaluation approach. However, COMET exhibits specific limitations that affect its practical applicability. Research by Amrhein and Sennrich (2022) and Glushkova et al. (2023) identified COMET's insensitivity to discrepancies in numbers and named entities, highlighting the need for targeted improvements in specific error detection capabilities. Hybrid and ensemble approaches have emerged as effective solutions to these limitations. Glushkova et al. (2023) demonstrated that combining traditional lexical metrics like BLEU with neural metrics such as COMET, improve the evaluation robustness by leveraging the strengths of both approaches. Extensions such as XCOMET (Guerreiro et al., 2023) also demonstrated strong correlation with human judgments, outperforming other pretrained metrics in most cases. These developments established COMET as a foundational framework for neural evaluation metric development.

Prism (Paraphrase-based Reference-free Evaluation) (Thompson & Post, 2020) introduced a novel approach using a multilingual neural machine translation model as an unbiased paraphraser for evaluation. This approach enables reference-free evaluation while maintaining robust correlation with human judgments across multiple languages. The metric's strength lies in its ability to provide system-level correlations that remain stable across diverse language pairs and domains. Prism's

multilingual training enables cross-lingual quality assessment without requiring reference translations, making it particularly valuable for low-resource language evaluation. However, research by Vamvas et al. (2023) revealed sensitivity to machine-generated references, indicating potential limitations when evaluating translations produced by systems similar to those used in Prism's training. This finding highlights the importance of considering training data composition in neural metric development.

GEMBA (GPT-based Evaluation Metric) (Kocmi & Federmann, 2023a, 2023b) represents a new type of neural evaluation metrics that leveraging large language models (LLMs), specifically GPT models (Radford et al., 2018), for MT evaluation. The metric employs carefully designed prompt templates to elicit quality judgments from language models. GEMBA demonstrates competitive performance in system ranking tasks and shows promise for document-level evaluation scenarios. Its flexibility in prompt design enables adaptation to different evaluation criteria and quality dimensions, providing versatility not available in more rigid metric frameworks. However, as noted by Bleiker (2023), the reliance on proprietary language models raises concerns about LLMs' proprietary nature and reproducibility, limiting the metric's applicability in research contexts that require open and reproducible evaluation procedures.

The WMT24 Metrics Shared Task (Freitag et al., 2024) provided comprehensive evaluation of current metric performance, revealing significant advances in neural approaches. The results (see **Table 1**) demonstrated that neural metrics were found to perform significantly better than lexical metrics, with BLEU, SPBLEU, and CHRF ranking 23rd, 22nd, and 20th, respectively, out of the 26 evaluated metrics. Top-performing metrics included MetricX-24-Hybrid (Juraska et al., 2024), XCOMET (Guerreiro et al., 2023), and MetaMetrics-MT (Anugraha et al., 2024), with MetaMetrics-MT, which is not an autonomous metric; instead, it harnesses multiple existing metrics, integrating them systematically, achieving an average correlation of 0.725. These findings underscore the effectiveness of hybrid methodologies.

Table 1: WMT24 Metrics Performance Comparison

Metric	Type	Correlation	Rank	Key Features
MetaMetrics-MT	Neural	0.725	1	Multi-dimensional analysis
MetricX-24-Hybrid	Neural	0.718	2	Error span features
XCOMET	Neural	0.695	3	Cross-lingual optimization
CHRF	Lexical	0.442	20	Character-level matching
SPBLEU	Lexical	0.435	22	Sentence-piece BLEU
BLEU	Lexical	0.428	23	N-gram matching

6. Limitations and Challenges

Despite significant advances in MT evaluation methodologies, current approaches face numerous limitations that continue to challenge researchers and practitioners. These limitations span multiple dimensions including fairness, robustness, interpretability, and practical applicability across diverse contexts and user populations. Research by Moghe et al. (2023) highlights several ways current MT metrics can be biased, leading to unfair evaluations. These biases can affect reliability and fairness, including:

1- Unfair evaluation of marginalized subpopulations due to inherent biases in data or metrics.

2-Sensitivity to paraphrases, leading to unfair penalization of translations with different but valid expressions.

3- Translationese, unnatural artifacts from the task language present in the test language during manual translation, overestimating performance.

4-Variable sensitivity to different MT errors, leading to biased evaluations based on critical errors for certain tasks.

5-Negligible correlation between segment-level scores and end task success/failure outcomes, indicating inability to accurately capture the quality of translations at a granular level, leading to biased evaluations that do not reflect the true performance of the translations in specific contexts.

Zhao et al. (2024) also highlight challenges in QE (Quality Estimation; reference-less metrics), including the lack of manually annotated data, especially for low-resource languages. Acquiring sufficient annotated data is costly and hinders QE research. Current QE approaches focus on

sentence-level tasks, with limited work on word-level and document-level QE. Word-level QE methods are few and often lack performance, but they can extract more fine-grained information.

Zouhar et al. (2024) highlight the challenges of maintaining accuracy and reliability across different domains, especially when fine-tuned on specific datasets. They introduced a comprehensive MQM annotated dataset for the biomedical domain, covering 11 language pairs. It includes translations and MQM annotations from 21 participants in the WMT21 biomedical translation shared task. Expert linguists re-translated the original biomedical test set, annotated, and corrected it. The dataset (comprising around 25,000 segment-level annotations) exhibited more Critical/Major errors compared to the WMT MQM dataset, underscoring the unique challenges of the biomedical domain and metric generalization across domains.

Research by Agrawal et al. (2024) found that most metrics exhibit high variance when evaluating HQ translations, indicating inconsistencies in scoring. Metrics like xCOMET-XXL and MetricX-23 showed higher correlations with human judgments in some settings but still fell short in others. GEMBA-MQM achieved the highest F1 score for detecting HQ-ZERO (high-quality translations that receive zero MQM scores from human evaluators) translations. However, Agrawal et al. (2024) noted a potential bias towards outputs generated by GPT-4, which could skew the evaluation results. Another limitation is the metrics' inability to distinguish between translations of the same source text, which is crucial for assessing subtle differences in translation quality. Agrawal et al. (2024) suggest using contrastive objectives or exposing the metrics to multiple translations to mitigate this issue, but it remains a challenge.

Automated metrics, despite their effectiveness in providing quantitative assessments, often lack insights into their scores, hindering the understanding and improvement of NLG (Natural Language Generation) systems (Leiter et al., 2022). Methods like SHAP (SHapley Additive exPlanations) enhance explainability by assigning importance scores to individual words in hypotheses and references (Leiter et al., 2022). Incorporating uncertainty into the assessment process also improves the reliability of quality predictions, as noted by Zerva et al. (2022).

7. Meta-Evaluation and Correlation Analysis in MT Evaluation

For over a decade, the Workshop on Machine Translation has conducted a shared task as a meta-evaluation platform for automatic metrics. Meta-evaluation quantifies a metric's performance by measuring the agreement between its scores and human-annotated scores on a substantial corpus of translations. Pearson's r and Spearman's ρ are commonly used correlation coefficients. Pearson's coefficient measures the linear relationship between two input vectors, calculated by dividing their covariance by the product of their variances. Spearman's coefficient is similar but applies to the ranks of the input variables (Deutsch et al., 2023).

Despite Pearson correlation's common use for ranking MT metrics, it has limitations, particularly its sensitivity to outliers, which can disproportionately affect evaluation results (Ma et al., 2019; Mathur et al., 2020). This issue arises because Pearson correlation doesn't account for the magnitude of differences between metric scores and human judgments, leading to misleading conclusions about system performance. Consequently, the outsized impact of outliers motivated the switch to pairwise accuracy in 2021 (Thompson et al., 2024).

Kendall's τ, a statistic estimating agreement between two sets of measurements based on their ranks, was used to measure the correlation between metrics and human judgments at the segment level before WMT23 (Perrella et al., 2024, p. 6). Like the Pearson coefficient, Kendall's τ has limitations. It's sensitive to noise in gold pairwise rankings, where pairs deemed not significantly different are often excluded (Freitag et al., 2022, p. 55). Deutsch et al. (2023, p. 4) also point out that the way existing variants of Kendall's τ handle ties (two values either concordant discordant) introduce blind spots in meta-evaluation and allow metrics to exploit τ-specific properties to improve correlations. This leads Kendall's τ to potentially penalize a metric for accurately predicting ties. To address this, Deutsch et al. (2023, p. 5) propose substituting Kendall's τ with a version of pairwise accuracy designed to handle ties.

Pairwise Accuracy (PA) compares the rankings produced by an automatic metric with those derived from human judgments. It determines which system provides better translation quality by focusing on the direction of differences rather than magnitude. PA uses a binary approach, scoring agreement between human and metric rankings as 1

(agreement) or 0 (disagreement) (Kocmi et al., 2021, pp. 4–5). This simplicity makes PA straightforward but has limitations. The binary nature disregards weak or statistically insignificant system preferences, leading to tied scores for multiple metrics, especially when evaluating many systems with close rankings (Thompson et al., 2024, pp. 2–3).

Thompson et al. (2024) introduce Soft Pairwise Accuracy (SPA), an improvement over traditional PA that addresses its limitations by incorporating statistical significance. SPA provides a more stable and reliable measure by considering both the accuracy of human pairwise rankings and their confidence. Instead of binarizing preferences, SPA uses continuous values between 0 and 1 to represent confidence. If a metric's preference between two systems is not statistically significant, SPA assigns a partial score reflecting uncertainty, unlike PA which assigns a definitive 1 or 0. It also penalizes metrics for showing unwarranted high confidence in cases of indifference or low significance, capturing a more detailed and statistically robust alignment between metrics and human judgments. These advantages have led to SPA being recognized as a more reliable for meta-evaluation, and paved the way for it to be implemented in the WMT24 Metrics Shared Task (Freitag et al., 2024).

8. Analysis of the Performance of Different Metrics under Different Systems

Al-Mahmood's (2025) latest MT evaluation dataset provides insights into the comparative performance of various metrics such as COMET, BLEU, chrF2, TER, and GEMBA-MQM. This dataset gives us an empirical demonstration of metric behavior, including its consistency, quality, and inter-metric relationships of these metrics. The dataset analyzes translation systems across two pairs: English to Arabic (Task 1) and Arabic to English (Task 2), and covers samples from commercial MT, human translators, and AI-based systems.

Upon inspecting the dataset, it became clear that BLEU, chrF2, and TER generally aligned with COMET scores (see **Figure 1**). In both tasks, the four metrics produced strong inter-metric agreement. TER, which measures the number of edits required to correct a translation (where lower is better), consistently showed the expected negative relationship with the other metrics. The temperature settings in AI systems also had minimal impact on translation outcomes ($|r| < 0.2$). These results were concerning given how they compare to GEMBA-MQM's error rates.

Figure 1: Pearson Correlation Heatmap of Automatic Metrics

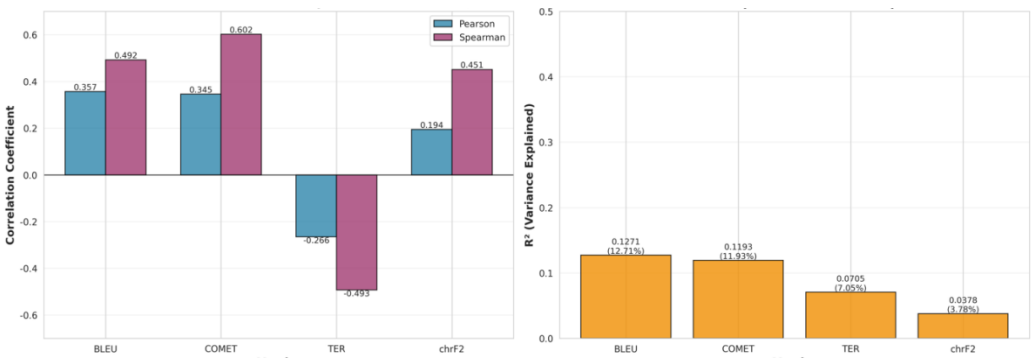
Critical, major, and minor errors captured by GEMBA-MQM provided a distinct perspective compared to other metrics, awarding higher scores to systems that exhibited poor performance under different metrics. For instance, in Task 1, OpenAI o1 achieved comparable low scores across COMET, BLEU, chrF2, and TER (62.76, 9.09, 28.65, and 96.38, respectively), while the system's GEMBA-MQM score was 92.6, with 3, 5, and 19 critical, major, and minor errors, respectively.

GEMBA-MQM demonstrated superior accuracy in identifying errors within human translations. For instance, in Task 2, HT2 achieved commendable performance on COMET, BLEU, chrF2, and TER (86.62, 36.84, 64.04, and 49.74, respectively). However, it yielded a significantly lower GEMBA-MQM score of 21.1, with 22, 78, and 179 errors categorized as critical, major, and minor, respectively. This underscores the significance of employing multi-metric evaluation and highlights the superiority of prompt-based metrics such as GEMBA-MQM over conventional metrics like BLEU, chrF2, TER, and restricted neural metrics like COMET.

The Pearson correlation analysis (see **Figure** 1), which examines linear relationships, showed relatively weak associations between GEMBA-

MQM and the automatic metrics. BLEU achieved r = 0.3566 (R^2 = 12.71%), COMET r = 0.3454 (R^2 = 11.93%), TER r = -0.2656 (R^2 = 7.05%), and chrF2 r = 0.1944 (R^2 = 3.78%). None of these correlations reached statistical significance. The low R^2 values suggest that, assuming linearity, these metrics explain only 4–13% of the variance in GEMBA-MQM scores. the low R^2 values (<13%) confirm that traditional metrics have limited predictive power, explaining only a small portion of GEMBA-MQM's variance. This finding highlights that GEMBA-MQM captures unique dimensions of translation quality not represented by conventional measures.

Figure 2: Comparison of Pearson and Spearman Coefficients (left), R^2 Values (right)



On the other hand, based on Spearman correlation coefficients (see **Figure 2**), which capture monotonic relationships, COMET showed the strongest association with GEMBA-MQM (ρ = 0.6018, $p < 0.01$). TER followed with a negative correlation of ρ = -0.4928 ($p < 0.05$), reflecting its inverse quality scale. BLEU (ρ = 0.4925, $p < 0.05$) and chrF2 (ρ = 0.4511, $p < 0.05$) both demonstrated moderate positive correlations. Spearman correlations indicated that all four metrics maintain statistically significant monotonic relationships with GEMBA-MQM. COMET exhibited the strongest correlation, followed by TER, BLEU, and chrF2. Although these four automatic metrics (BLEU, COMET, chrF2, TER) are highly inter-correlated ($r > 0.84$), their only moderate correlations with GEMBA-MQM indicate that GEMBA-MQM serves as a more independent quality assessment.

The dataset (Al-Mahmood, 2025) also revealed important insight about conducting human evaluation with limited number of evaluators. When inter-rater reliability was evaluated using Krippendorff's Alpha (see **Table 2**), a robust metric suitable for assessing agreement in the presence of incomplete data or varying measurement levels, the values

obtained for both accuracy and fluency ratings across two distinct evaluation tasks revealed generally low levels of agreement among the three human raters. For the first evaluation task, Krippendorff's Alpha for accuracy ratings was 0.34, while for fluency ratings, it was 0.336. In the second evaluation task, the reliability scores were notably lower, with Krippendorff's Alpha for accuracy at 0.044 and for fluency at 0.043.

Table 2: Unweighted Agreement Coefficients of Accuracy and Fluency Ratings

Task	Rating	Method	Coeff	StdErr	95% C.I.	P-Value
1	Accuracy	Krippendorff's Alpha	0.340	0.047	(0.247,0.433)	5.210e-11
1	Accuracy	Percent Agreement	0.577	0.035	(0.508,0.646)	0.000e+00
1	Fluency	Krippendorff's Alpha	0.336	0.047	(0.242,0.431)	9.792e-11
1	Fluency	Percent Agreement	0.577	0.035	(0.508,0.646)	0.000e+00
2	Accuracy	Krippendorff's Alpha	0.308	0.044	(0.22,0.395)	1.739e-10
2	Accuracy	Percent Agreement	0.580	0.033	(0.514,0.646)	0.000e+00
2	Fluency	Krippendorff's Alpha	0.274	0.043	(0.189,0.36)	3.446e-09
2	Fluency	Percent Agreement	0.560	0.033	(0.495,0.625)	0.000e+00

The consistently low Krippendorff's Alpha values indicate substantial disagreement among the three human raters in their assessments of both accuracy and fluency. Several factors may contribute to such low inter-rater agreement, including ambiguous rating guidelines, subjective interpretation of evaluation criteria, or inadequate rater training. The particularly low scores in the second task suggest that the rating criteria or evaluation materials were even more challenging to apply consistently in that context.

These findings underscore the critical importance of establishing robust evaluation protocols for human assessment. When inter-rater reliability is low, individual ratings become less reproducible across different

raters, limiting the validity and generalizability of the evaluation results. This highlights a fundamental challenge: without clear, well-defined criteria and adequate rater calibration, human evaluations may introduce more noise than signal into the assessment process. In such circumstances, complementary automatic metrics can provide more consistent and reproducible measurements, though they should ideally be used alongside—rather than as replacements for—improved human evaluation protocols.

9. Conclusion

Machine translation evaluation has undergone remarkable transformation, reflecting both the advancement of translation systems and the growing understanding of translation quality as a multifaceted phenomenon requiring nuanced assessment approaches. This review reveals that each evaluation paradigm offers distinct strengths while simultaneously presenting inherent limitations that constrain comprehensive quality assessment.

Human evaluation methods, including Direct Assessment (DA), Multidimensional Quality Metrics (MQM), and Error Span Annotation (ESA), remain the gold standard for capturing nuanced semantic, pragmatic, and stylistic dimensions of translation quality. However, matters regarding evaluator training, scale, and cost limit their applicability in resource-constrained experiments. While basic adequacy and fluency measurements provide valuable insights, they fall short in objectivity and consistency, as demonstrated through inter-rater reliability analysis of Al-Mahmood's (2025) English⇔Arabic dataset, with Krippendorff's Alpha values ranging from 0.044 to 0.340, indicating substantial variability in human judgment.

Automatic lexical metrics such as BLEU, METEOR, and TER revolutionized the field by enabling rapid, large-scale comparative evaluations essential for iterative system development. Despite their computational efficiency and reproducibility, these metrics demonstrate fundamental limitations in capturing semantic nuances and contextual appropriateness. The empirical analysis revealed that while BLEU, chrF2, and TER maintain strong inter-correlations ($r > 0.84$), they explain only a fraction of the variance in advanced neural metrics like GEMBA-MQM (4–13%), highlighting their inability to capture the multidimensional nature of translation quality.

GEMBA-MQM demonstrated superior diagnostic capabilities, particularly in identifying errors within human translations that other metrics failed to detect, despite its lack of interpretability and black-box nature. The weak to moderate correlations between GEMBA-MQM and traditional metrics (Pearson's *r* ranging from 0.19 to 0.36) suggest that prompt-based neural metrics capture unique dimensions of translation quality not represented by conventional measures. This finding reinforces the necessity of multi-metric evaluation strategies that leverage the complementary strengths of different approaches rather than relying on any single metric.

The evolution of meta-evaluation techniques from Pearson, Spearman, and Kendall's correlations to more robust measures like Soft Pairwise Accuracy (SPA) reflects the field's ongoing efforts to improve the reliability and statistical soundness of metric assessment. SPA's incorporation of statistical significance and confidence levels addresses critical limitations of earlier approaches, providing more stable and nuanced measures of metric-human agreement. This methodological advancement has important implications for future benchmark development and metric validation procedures.

Future Directions and Recommendations

Several critical areas require attention to advance MT evaluation methodologies. First, the need for explainable evaluation methods becomes increasingly critical as MT evaluation systems are deployed in high-stakes applications. Future research must balance the sophistication of neural approaches with the interpretability requirements of practical applications. Promising directions include development of metrics that provide fine-grained error analysis, uncertainty quantification, and detailed justification for quality judgments. These capabilities would support both diagnostic system improvement and transparent decision-making in translation deployment scenarios.

Second, the computational requirements of current neural metrics limit their accessibility and practical applicability, particularly for resource-constrained organizations. Future research should investigate efficient metric architectures that maintain high performance while reducing computational demands. User-configurable evaluation systems that allow stakeholders to specify quality priorities and adapt metric behavior to specific requirements could bridge the gap between general-purpose metrics and specialized evaluation needs, enabling more targeted and

relevant quality assessment while maintaining systematic evaluation procedures.

Third, the challenge of evaluating high-quality translations, where most metrics exhibit high variance and inconsistent scoring, represents another critical area requiring attention. As MT systems increasingly produce near-human quality outputs, evaluation methodologies must evolve to distinguish subtle differences in translation quality that have significant implications for user experience and task success.

Notes

[1]This paper is based on the MA thesis titled "Investigating the Variation of Competence in Human and Machine Translation: A Benchmark Study", conducted by the first researcher, Hussein Ali Khudhair Al-Mahmood, under the supervision of the second researcher, Asst. Prof. Abdulsalam Abdulmajeed Al-Ogaili.

References

Agrawal, S., Farinhas, A., Rei, R., & Martins, A. F. T. (2024). *Can Automatic Metrics Assess High-Quality Translations?* (No. arXiv:2405.18348). arXiv. https://doi.org/10.48550/arXiv.2405.18348

Al-Mahmood, H. (2025). *Evaluation of LLM-based MT against Human Translators* [Dataset]. Zenodo. https://doi.org/10.5281/ZENODO.17156571

Amrhein, C., & Sennrich, R. (2022). *Identifying Weaknesses in Machine Translation Metrics Through Minimum Bayes Risk Decoding: A Case Study for COMET* (No. arXiv:2202.05148). arXiv. https://doi.org/10.48550/arXiv.2202.05148

Anugraha, D., Kuwanto, G., Susanto, L., Wijaya, D. T., & Winata, G. (2024). MetaMetrics-MT: Tuning Meta-Metrics for Machine Translation via Human Preference Calibration. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 459–469). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.32

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, & C. Voss (Eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72). Association for Computational Linguistics. https://aclanthology.org/W05-0909

Birch, A., Abend, O., Bojar, O., & Haddow, B. (2016). HUME: Human UCCA-Based Evaluation of Machine Translation. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1264–1274). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1134

Bleiker, N. (2023). *Evaluation of Pre-trained Metrics and ChatGPT as Document-level Machine Translation Metrics* [Master's Thesis, University of Zurich]. https://www.zora.uzh.ch/entities/publication/51583e20-2039-49fa-8813-947ab86d812b

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). *Approaches to Human and Machine Translation Quality Assessment: From Principles to Practice* (pp. 9–38). https://doi.org/10.1007/978-3-319-91241-7_2

Daelemans, W., & Hoste, V. (Eds.). (2009). *Evaluation of translation technology*. Artesis Univ. College Antwerp, Department of Translators & Interpreters.

Deutsch, D., Foster, G., & Freitag, M. (2023). *Ties Matter: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration* (No. arXiv:2305.14324). arXiv. https://doi.org/10.48550/arXiv.2305.14324

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805; Version 1). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Duh, K. (2008). Ranking vs. Regression in Machine Translation Evaluation. In C. Callison-Burch, P. Koehn, C. Monz, J. Schroeder, & C. S. Fordyce (Eds.), *Proceedings of the Third Workshop on Statistical Machine Translation* (pp. 191–194). Association for Computational Linguistics. https://aclanthology.org/W08-0331

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, *9*, 1460–1474. https://doi.org/10.1162/tacl_a_00437

Freitag, M., Mathur, N., Deutsch, D., Lo, C.-K., Avramidis, E., Rei, R., Thompson, B., Blain, F., Kocmi, T., Wang, J., Adelani, D. I., Buchicchio, M., Zerva, C., & Lavie, A. (2024). Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 47–81). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.2

Freitag, M., Rei, R., Mathur, N., Lo, C., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., & Martins, A. F. T. (2022). Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, … M. Zampieri (Eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 46–68). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.2/

Giménez, J., & Màrquez, L. (2007). Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In C. Callison-Burch, P. Koehn, C. S. Fordyce, & C. Monz (Eds.), *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 256–264). Association for Computational Linguistics. https://aclanthology.org/W07-0738/

Glushkova, T., Zerva, C., & Martins, A. F. T. (2023). *BLEU Meets COMET: Combining Lexical and Neural Metrics Towards Robust Machine Translation Evaluation* (No. arXiv:2305.19144). arXiv. http://arxiv.org/abs/2305.19144

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013). Continuous Measurement Scales in Human Evaluation of Machine Translation. In A. Pareja-Lora, M. Liakata, & S. Dipper (Eds.), *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 33–41). Association for Computational Linguistics. https://aclanthology.org/W13-2305

Guerreiro, N. M., Rei, R., Stigt, D. van, Coheur, L., Colombo, P., & Martins, A. F. T. (2023). *xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection* (No. arXiv:2310.10482). arXiv. https://doi.org/10.48550/arXiv.2310.10482

House, J. (1977). A Model for Assessing Translation Quality. *Meta: Journal des traducteurs*, *22*(2), 103. https://doi.org/10.7202/003140ar

Juraska, J., Deutsch, D., Finkelstein, M., & Freitag, M. (2024). MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 492–504). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.35

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., … Zouhar, V. (2024). Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 1–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.1

Kocmi, T., & Federmann, C. (2023a). *GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4* (No. arXiv:2310.13988). arXiv. https://doi.org/10.48550/arXiv.2310.13988

Kocmi, T., & Federmann, C. (2023b). *Large Language Models Are State-of-the-Art Evaluators of Translation Quality* (No. arXiv:2302.14520; Version 2). arXiv. https://doi.org/10.48550/arXiv.2302.14520

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., & Menezes, A. (2021). *To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation* (No. arXiv:2107.10821). arXiv. https://doi.org/10.48550/arXiv.2107.10821

Kocmi, T., Zouhar, V., Avramidis, E., Grundkiewicz, R., Karpinska, M., Popović, M., Sachan, M., & Shmatova, M. (2024). *Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation* (No. arXiv:2406.11580). arXiv. https://doi.org/10.48550/arXiv.2406.11580

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Koehn, P. (2020). *Neural Machine Translation* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108608480

Koehn, P., & Monz, C. (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. In P. Koehn & C. Monz (Eds.), *Proceedings on the Workshop on Statistical Machine Translation* (pp. 102–121). Association for Computational Linguistics. https://aclanthology.org/W06-3114

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A Survey on Evaluation Metrics for Machine Translation. *Mathematics*, *11*(4), Article 4. https://doi.org/10.3390/math11041006

Leiter, C., Lertvittayakumjorn, P., Fomicheva, M., Zhao, W., Gao, Y., & Eger, S. (2022). *Towards Explainable Evaluation Metrics for Natural Language Generation* (No. arXiv:2203.11131). arXiv. https://doi.org/10.48550/arXiv.2203.11131

Li, H., Shi, T., Shang, Z., Han, Y., Zhao, X., Wang, H., Qian, Y., Qian, Z., Xu, L., Wu, M., Lyu, C., Wang, L., Tang, G., Luo, W., Xu, Z., & Zhang, K. (2025). *TransBench: Benchmarking Machine Translation for Industrial-Scale Applications* (No. arXiv:2505.14244). arXiv. https://doi.org/10.48550/arXiv.2505.14244

Li, S., Wang, J., Ali, F. D. M. A., Cherry, C., Deutsch, D., Briakou, E., Sousa-Silva, R., Cardoso, H. L., Stenetorp, P., & Adelani, D. I. (2025). *SSA-COMET: Do LLMs Outperform Learned Metrics in Evaluating MT for Under-Resourced African Languages?* (No. arXiv:2506.04557). arXiv. https://doi.org/10.48550/arXiv.2506.04557

Lommel, A. R., Burchardt, A., & Uszkoreit, H. (2013, November 28). Multidimensional quality metrics: A flexible system for assessing translation quality. *Proceedings of Translating and the Computer 35*. TC 2013, London, UK. https://aclanthology.org/2013.tc-1.6

Ma, Q., Wei, J., Bojar, O., & Graham, Y. (2019). Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 62–90). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5302

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4984–4997). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.448

Moghe, N., Sherborne, T., Steedman, M., & Birch, A. (2023). *Extrinsic Evaluation of Machine Translation Metrics* (No. arXiv:2212.10297). arXiv. https://doi.org/10.48550/arXiv.2212.10297

Musa, Z. Y., & Al-Maryani, J. K. (2021). Assessing the simultaneous interpreting outputs of trainee interpreters in Iraqi Departments of Translation. *Adab Al-Basrah Journal*, *95*(2), 19–66. https://www.basradab.edu.iq/wp-content/uploads/2024/09/16-زكريا-جاسم-خليفةE.pdf

Na'ja, S., & Abu-Mighnim, J. (2012). Analysis of morphological errors of Arabic language learners in light of the linguistic convergence. *Journal of Tikrit University for the Humanities*, *19*, 167–219.

National Research Council. (1966). *Language and Machines: Computers in Translation and Linguistics*. https://doi.org/10.17226/9547

Olive, J., Christianson, C., & McCary, J. (Eds.). (2011). *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer. https://doi.org/10.1007/978-1-4419-7713-7

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin

(Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Perrella, S., Proietti, L., Scirè, A., Barba, E., & Navigli, R. (2024). *Guardians of the Machine Translation Meta-Evaluation: Sentinel Metrics Fall In!* (No. 2408.13831v1). arXiv. https://arxiv.org/abs/2408.13831v1

Quah, C. K. (2006). *Translation and Technology*. Palgrave Macmillan UK. https://doi.org/10.1057/9780230287105

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Riccardi, A. (1999). *Attuali metodi di valutazione presso la SSLMIT*. SSLMIT. http://hdl.handle.net/11368/1688328

Sellam, T., Das, D., & Parikh, A. P. (2020). *BLEURT: Learning Robust Metrics for Text Generation* (No. 2004.04696v5). arXiv. https://arxiv.org/abs/2004.04696v5

Silva, B., Buchicchio, M., Van Stigt, D., Stewart, C., Moniz, H., & Lavie, A. (2024). Data-driven Asian Adapted MQM Typology and Automation in Translation Quality Workflows. *The Journal of Specialised Translation*, *41*, 98–126. https://doi.org/10.26034/cm.jostrans.2024.4713

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231. https://aclanthology.org/2006.amta-papers.25

Snover, M., Madnani, N., Dorr, B., & Schwartz, R. (2009). Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In C. Callison-Burch, P. Koehn, C. Monz, & J. Schroeder (Eds.), *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 259–268). Association for Computational Linguistics. https://aclanthology.org/W09-0441

Song, Y., Riley, P., Deutsch, D., & Freitag, M. (2025). *Enhancing Human Evaluation in Machine Translation with Comparative Judgment* (No. arXiv:2502.17797). arXiv. https://doi.org/10.48550/arXiv.2502.17797

Sparck Jones, K., & Galliers, J. R. (1995). *Evaluating natural language processing systems: An analysis and review*. Springer.

Thompson, B., Mathur, N., Deutsch, D., & Khayrallah, H. (2024). *Improving Statistical Significance in Human Evaluation of Automatic Metrics via Soft Pairwise Accuracy* (No. arXiv:2409.09598). arXiv. https://doi.org/10.48550/arXiv.2409.09598

Thompson, B., & Post, M. (2020). Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 90–121). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.8

Vamvas, J., Domhan, T., Trenous, S., Sennrich, R., & Hasler, E. (2023). *Trained MT Metrics Learn to Cope with Machine-translated References* (No. arXiv:2312.00536). arXiv. https://doi.org/10.48550/arXiv.2312.00536

Vilar, D., Leusch, G., Ney, H., & Banchs, R. E. (2007). Human Evaluation of Machine Translation Through Binary System Comparisons. In C. Callison-Burch, P. Koehn, C. S. Fordyce, & C. Monz (Eds.), *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 96–103). Association for Computational Linguistics. https://aclanthology.org/W07-0713

White, J. S. (2003). How to evaluate machine translation. In H. Somers (Ed.), *Computers and Translation: A translator's guide* (pp. 211–244). John Benjamins Publishing Company. https://doi.org/10.1075/btl.35.16whi

White, J. S., & O'Connell, T. A. (1996, October 2). Adaptation of the DARPA machine translation evlauation paradigm to end-to-end systems. *Conference of the Association for Machine Translation in the Americas*. AMTA 1996, Montreal, Canada. https://aclanthology.org/1996.amta-1.11

Yan, Y., Wang, T., Zhao, C., Huang, S., Chen, J., & Wang, M. (2023). BLEURT Has Universal Translations: An Analysis of Automatic Metrics by Minimum Risk Training. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5428–5443). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.297

Zerva, C., Glushkova, T., Rei, R., & Martins, A. F. T. (2022). *Disentangling Uncertainty in Machine Translation Evaluation* (No. arXiv:2204.06546). arXiv. https://doi.org/10.48550/arXiv.2204.06546

Zhao, H., Liu, Y., Tao, S., Meng, W., Chen, Y., Geng, X., Su, C., Zhang, M., & Yang, H. (2024). *From Handcrafted Features to LLMs: A Brief Survey for Machine Translation Quality Estimation* (No. 2403.14118v1). arXiv. https://arxiv.org/abs/2403.14118v1

Zouhar, V., Ding, S., Currey, A., Badeka, T., Wang, J., & Thompson, B. (2024). *Fine-Tuned Machine Translation Metrics Struggle in Unseen Domains* (No. 2402.18747v2). arXiv. https://arxiv.org/abs/2402.18747v2