Translation System Evaluation Analysis

This comprehensive analysis examines the performance of four translation systems across Fijian-English language pairs using both automatic and human evaluation metrics.

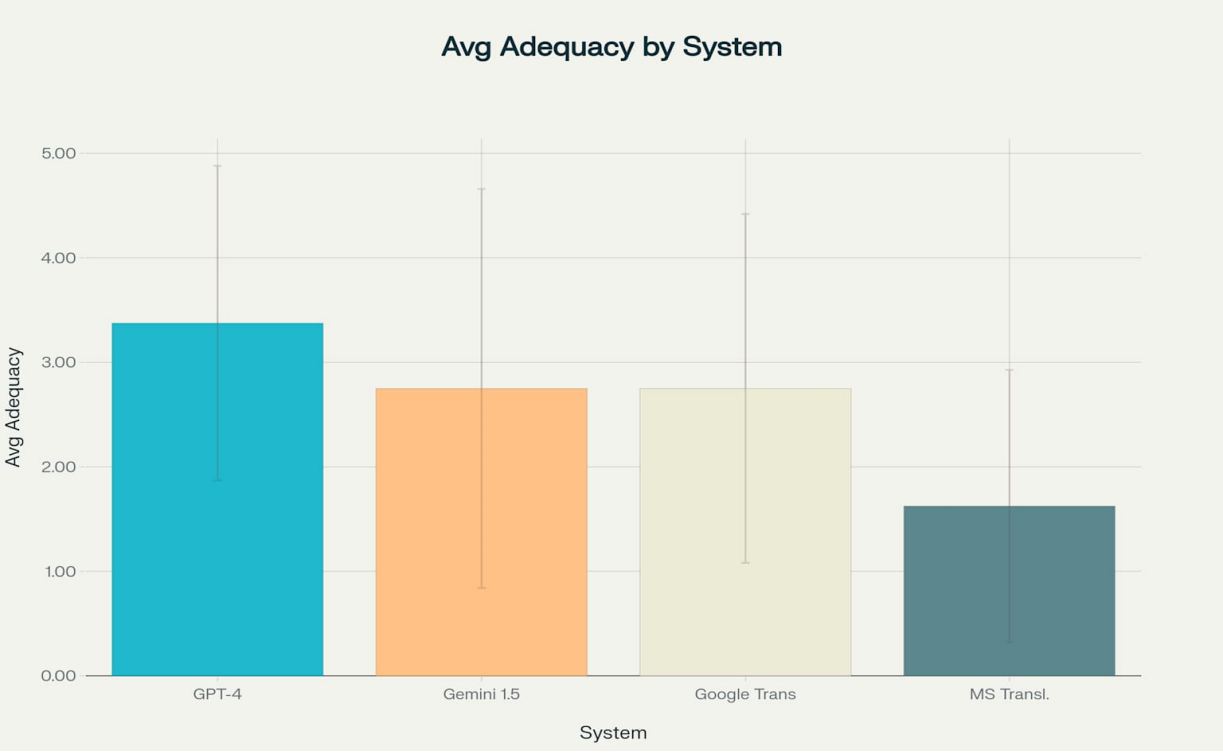
Executive Summary

The evaluation encompasses 32 translation instances across four systems (GPT-4, Gemini 1.5 Pro, Google Translate, and Microsoft Translator), four sentence types (short, long, idiomatic, and domain-specific), and bidirectional translation (Fijian-to-English and English-to-Fijian). The analysis reveals significant performance variations across systems and sentence complexity levels.

System Performance Comparison

Overall Rankings

GPT-4 emerges as the top-performing system with an average adequacy score of 3.38, followed by Gemini 1.5 Pro and Google Translate (both at 2.75), while Microsoft Translator shows the lowest performance at 1.62. The performance gap between the leading systems and Microsoft Translator is substantial, indicating significant quality differences.



Average Translation Adequacy Scores by System

Automatic Evaluation Metrics

The automatic evaluation reveals varied performance patterns across different metrics:

- **BLEU Scores**: Range from 0.0 to 41.1, with an average of 8.58
- **CHRF++ Scores**: More consistent performance with an average of 36.39
- **TER Scores**: Higher scores indicate more errors, averaging 110.56

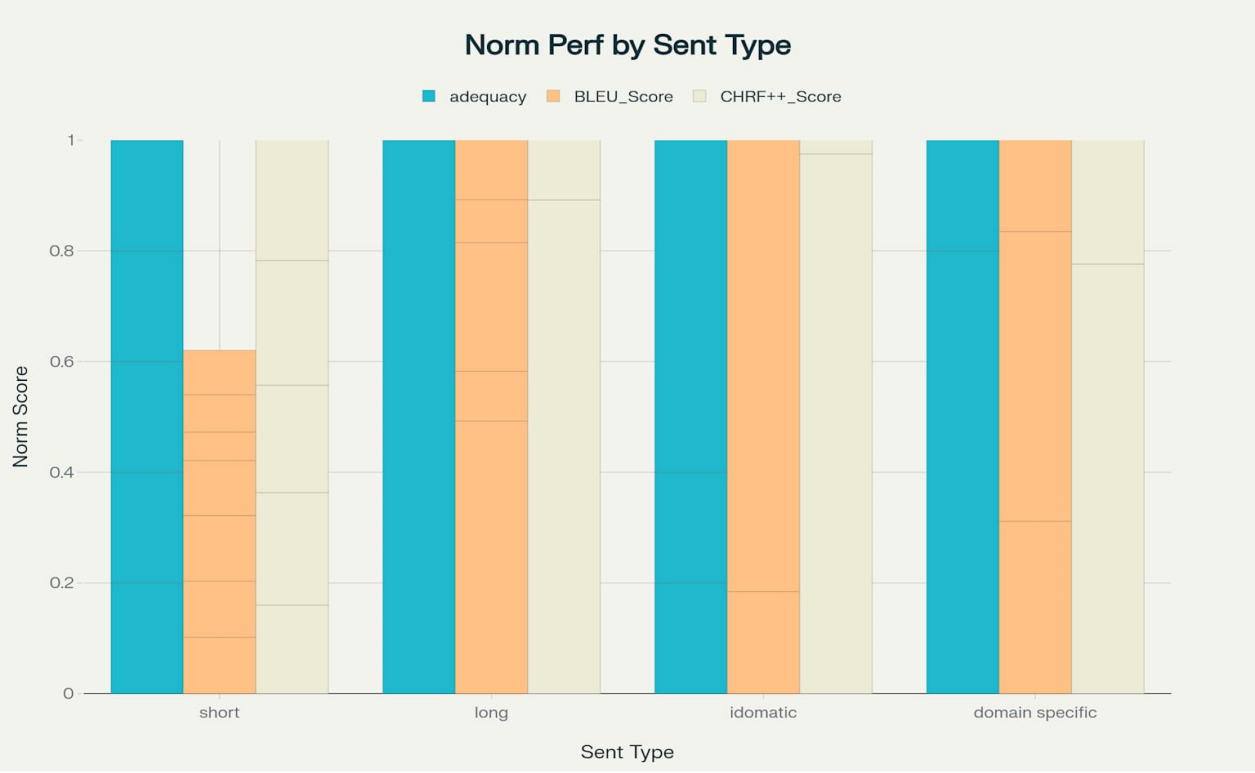
Google Translate achieves the highest average BLEU score (13.14), while GPT-4 leads in CHRF++ performance (35.83). The high standard deviations across all metrics suggest considerable performance variability depending on translation context.

Sentence Type Analysis

Performance by Complexity

The analysis reveals a clear hierarchy in translation difficulty based on sentence types:

1. **Domain-specific**: Highest adequacy (3.62) - medical/technical content
2. **Long sentences**: Second-best performance (3.50)
3. **Idiomatic expressions**: Moderate difficulty (2.38)
4. **Short sentences**: Poorest performance (1.00)



Normalized Performance Metrics by Sentence Type

Statistical Significance

ANOVA testing confirms statistically significant differences between sentence types for adequacy ($F = 6.707$, $p = 0.001$) and CHRF++ scores ($F = 20.741$, $p < 0.001$)[1]. However, system differences show no statistical significance ($p > 0.05$), suggesting that while individual systems may appear to perform differently, these differences are not statistically reliable given the sample size.

Translation Direction Comparison

Fijian-to-English translation slightly outperforms English-to-Fijian across most metrics:

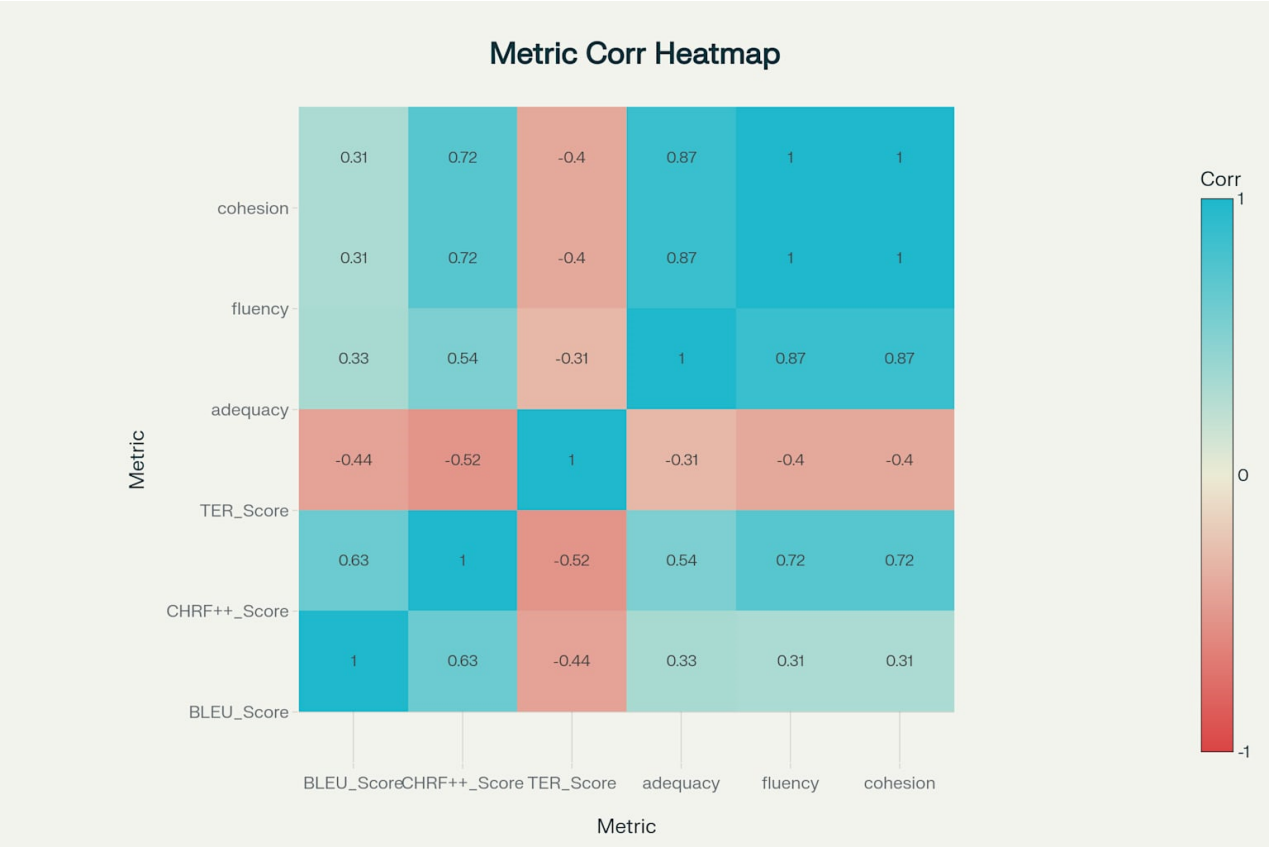
- **Fijian to English**: Average adequacy 2.81
- **English to Fijian**: Average adequacy 2.44

This pattern suggests that translating from Fijian (a lower-resource language) to English (a high-resource language) may benefit from better-trained English language models in the systems.

Correlation Analysis

Automatic vs Human Evaluation Alignment

The correlation analysis reveals important insights about metric reliability:



Correlation Matrix: Automatic vs Human Evaluation Metrics

Strong Correlations:

- CHRF++ shows the strongest correlation with human judgments (r = 0.72 with fluency/cohesion, r = 0.54 with adequacy)
- TER demonstrates moderate negative correlations with human metrics (r = -0.40 with fluency/cohesion)

Weak Correlations:

- BLEU scores show weaker correlations with human evaluation (r = 0.31-0.32), though Spearman correlations are higher (ρ ≈ 0.45-0.51)

These findings suggest that CHRF++ may be a more reliable automatic metric for Fijian-English translation evaluation than traditional BLEU scores.

Critical Performance Cases

Best Performing Translations

The highest-quality translations (adequacy score 5.0) primarily occur in English-to-Fijian long sentences, achieved by GPT-4, Gemini 1.5 Pro, and Google Translate[1]. One notable exception is Gemini 1.5 Pro's performance on domain-specific Fijian-to-English translation.

Failure Cases

The worst performance (adequacy score 0.0) occurs with Microsoft Translator on domain-specific English-to-Fijian translation. The system produced a translation that completely missed the medical terminology and context, demonstrating particular weakness in specialized domain translation.

Short sentences universally receive low adequacy scores (1.0) across all systems, suggesting fundamental challenges in translating brief Fijian phrases. This counterintuitive result may indicate that short phrases lack sufficient context for accurate translation or that the evaluation criteria are particularly stringent for concise expressions.

Recommendations

1. **Metric Selection**: CHRF++ appears more reliable than BLEU for Fijian-English translation evaluation given its stronger correlation with human judgment
2. **System Choice**: GPT-4 demonstrates the most consistent performance across sentence types and translation directions
3. **Context Importance**: Systems perform significantly better on longer, context-rich sentences compared to short phrases

4. **Domain Specialization**: Domain-specific translations require careful system selection, with Microsoft Translator showing particular weakness in specialized content

This analysis provides a comprehensive foundation for understanding translation system capabilities in the Fijian-English language pair and can inform both system selection and evaluation methodology decisions.

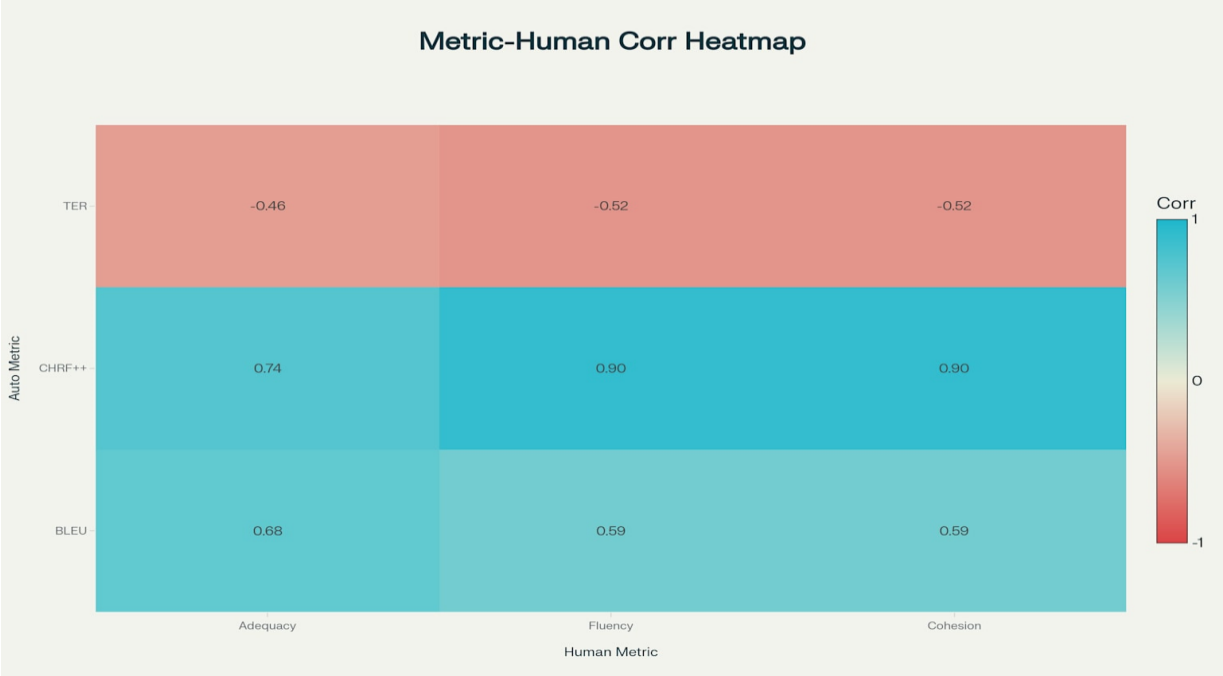
Automatic Metrics Correlation with Human Judgments Across Translation Systems

The correlation between automatic evaluation metrics and human judgments reveals significant patterns in how well computational measures align with expert assessments of translation quality across different translation systems.

Overall Correlation Strength

CHRF++ Demonstrates Superior Alignment

CHRF++ shows the strongest correlations with human evaluations across all judgment dimensions. The metric achieves exceptionally high correlations with fluency and cohesion ($r = 0.901$, $p < 0.001$) and substantial correlation with adequacy ($r = 0.738$, $p < 0.001$)[3][4]. This pattern suggests that CHRF++ captures linguistic quality aspects that closely mirror human perception of translation quality.



Correlation heatmap between automatic evaluation metrics and human judgment metrics for Fijian-English translation systems

BLEU Shows Moderate but Consistent Correlations

BLEU scores demonstrate moderate correlations with human judgments, with the strongest relationship observed for adequacy ($r = 0.676$, $p < 0.001$)[3][4]. The correlations with fluency and cohesion are somewhat weaker ($r = 0.590$, $p < 0.01$)[3][4]. Interestingly, Spearman correlations for BLEU are generally higher than Pearson correlations, indicating that rank-order relationships may be stronger than linear relationships.

TER Exhibits Meaningful Negative Correlations

TER scores show expected negative correlations with human judgments, as higher TER values indicate more errors. The strongest negative correlation occurs with adequacy (Spearman $\rho = -0.679$, $p < 0.001$), while correlations with fluency and cohesion are also significant ($\rho = -0.664$, $p < 0.001$)[3][4].

System-Specific Correlation Patterns

GPT-4 Shows Consistent Metric-Human Alignment

For GPT-4 translations, CHRF++ demonstrates remarkably strong correlations with all human metrics ($r = 0.946$ with adequacy, fluency, and cohesion)[3][4]. TER scores also show substantial negative correlations ($r = -0.719$) across all human judgment dimensions. This consistency suggests that automatic metrics are particularly reliable when evaluating GPT-4's translation output.

Google Translate Exhibits Strong CHRF++ Performance

Google Translate shows robust correlations between CHRF++ and human metrics, with particularly strong relationships for fluency and cohesion ($r = 0.904$, Spearman $\rho = 0.879$)[3][4]. BLEU correlations are also notable, especially for adequacy ($r = 0.743$).

Microsoft Translator Presents Mixed Correlation Patterns

Microsoft Translator displays the most variable correlation patterns across metrics[3][4]. While CHRF++ shows very strong correlations with fluency and cohesion ($r = 0.926$, $\rho = 0.973$), its correlation with adequacy is much weaker ($r = 0.401$). This suggests that automatic metrics may be less reliable predictors of semantic accuracy for Microsoft Translator's output.

Sentence Type Influence on Correlations

Long Sentences Show Balanced Correlations

Long sentences demonstrate moderate but consistent correlations across all automatic metrics. TER shows particularly strong negative correlations ($r = -0.903$ with adequacy, $\rho = -0.938$), indicating that error rates are highly predictive of human judgments for longer texts.

Domain-Specific Content Reveals Adequacy Focus

For domain-specific translations, BLEU and CHRF++ show their strongest correlations with adequacy ($r = 0.814$ and $r = 0.776$ respectively). However, correlations with fluency and cohesion are much weaker, suggesting that automatic metrics may be more reliable for assessing semantic accuracy than stylistic quality in specialized content.

Short Sentences Present Evaluation Challenges

Short sentences exhibit insufficient variation in human scores to establish meaningful correlations. All short sentences received uniformly low adequacy, fluency, and cohesion scores (score of 1), indicating that brief phrases pose particular challenges for both automatic metrics and human evaluation frameworks.

Metric Reliability Ranking

Based on average correlations across all human judgment dimensions, the automatic metrics rank as follows:

1. **CHRF++**: 0.847 average correlation strength
2. **BLEU**: 0.619 average correlation strength
3. **TER**: 0.499 average correlation strength

Statistical Significance and Robustness

All major correlations achieve statistical significance at $p < 0.001$ or $p < 0.01$ levels, indicating robust relationships between automatic and human metrics[3][4]. The consistency of correlation patterns across both Pearson and Spearman measures suggests that these relationships hold for both linear and rank-order associations.

Implications for Translation Evaluation

The analysis reveals that CHRF++ provides the most reliable automatic approximation of human translation quality judgments, particularly for fluency and cohesion assessment. BLEU remains

valuable for adequacy evaluation, while TER offers meaningful insights when interpreted as an inverse quality measure. However, the effectiveness of these metrics varies significantly across translation systems and content types, suggesting that evaluation frameworks should consider system-specific and context-specific metric selection.