# Literature Survey: Automatic Evaluation Metrics for Statistical Machine Translation

**Topic:** Automatic Evaluation of Machine Translation Quality **Focus:** Comparison of BLEU, METEOR, TER, chrF, BEER, and COMET/BLEURT

---


## Abstract

Automatic evaluation of machine translation (MT) quality is crucial for system development and comparison. This survey reviews the major automatic MT evaluation metrics, analyzing their methodologies, strengths, weaknesses, and correlation with human judgments. We cover traditional n-gram-based metrics (BLEU, METEOR, TER), character-based metrics (chrF), embedding-based metrics (BEER), and neural metrics (COMET, BLEURT). Despite the emergence of neural approaches, BLEU remains the de facto standard due to its simplicity and interpretability, though it has well-documented limitations.


---

## Table of Contents


1. Introduction
2. BLEU: Bilingual Evaluation Understudy
3. METEOR: Metric for Evaluation of Translation with Explicit ORdering
4. TER: Translation Error Rate
5. chrF: Character n-gram F-score
6. BEER: Better Evaluation as Ranking
7. COMET and BLEURT: Neural Evaluation Metrics
8. Comparative Analysis
9. Why BLEU Persists
10. Conclusion
11. References


---

## 1. Introduction

### The Need for Automatic Evaluation

Machine translation evaluation traditionally relied on human judgments of adequacy and fluency [@callison2006re]. However, human evaluation is: - **Expensive**: Requires trained annotators - **Time-consuming**: Slow turnaround for system development - **Inconsistent**: Inter-annotator disagreement - **Not scalable**: Cannot evaluate all system variants

Automatic metrics address these issues by providing fast, cheap, and reproducible evaluation.

### Desirable Metric Properties

An ideal MT evaluation metric should: 1. **Correlate highly with human judgments** (primary goal) 2. **Be language-independent** or easily adaptable 3. **Be fast to compute** (enable rapid iteration) 4. **Be interpretable** (explain what it measures) 5. **Be robust** (consistent across domains) 6. **Have low variance** (stable across test sets)

No single metric satisfies all criteria, leading to a diverse landscape of evaluation approaches.

---

## 2. BLEU: Bilingual Evaluation Understudy

### 2.1 Methodology

BLEU [@papineni2002bleu] measures n-gram overlap between candidate translation and reference(s):

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where: - $p_n$ = modified n-gram precision (with clipping) - $w_n$ = uniform weights (typically 1/4 for n=1..4) - BP = $\min(1, e^{1-r/c})$ = brevity penalty

**Key Innovation:** Modified precision clips candidate n-gram counts by maximum reference counts, preventing gaming the metric with repetition.

### 2.2 Strengths

1. **Simplicity**: Easy to understand and implement
2. **Speed**: Very fast computation (O(n) in sentence length)
3. **Ubiquity**: Universally reported, enabling cross-study comparison
4. **Language-independence**: Requires no linguistic resources
5. **Correlation**: Reasonable corpus-level correlation with human judgments (0.6-0.7 Pearson's r)

### 2.3 Weaknesses

1. **Word-level only**: Ignores synonyms, paraphrases [@callison2006re]
2. **Recall blindness**: Only measures precision, not recall
3. **Sentence-level poor correlation**: Low correlation at sentence level (0.3-0.4)
4. **Word order insensitivity**: Bigrams capture only local reordering

5. **Morphology-insensitive**: Treats “run”, “running”, “ran” as completely different
6. **Reference dependency**: Requires multiple references for reliability
7. **Saturation**: Difficult to improve beyond ~40-50 BLEU

**Example of Limitation:** - Candidate: “the cat sat on the mat” - Reference: “the feline was sitting on the rug” - BLEU: Very low (different words), but semantically equivalent

### 2.4 Variants

- **BLEU-1, BLEU-2, BLEU-3, BLEU-4**: Individual n-gram orders
- **NIST**: Weighted n-grams by informativeness [@doddington2002automatic]
- **GLEU**: Google BLEU with different smoothing [@wu2016google]

---

## 3. METEOR: Metric for Evaluation of Translation with Explicit ORdering

### 3.1 Methodology

METEOR [@banerjee2005meteor; @denkowski2014meteor] improves on BLEU by incorporating: 1. **Stem matching**: “running” matches “run” 2. **Synonym matching**: Using WordNet/paraphrase tables 3. **Paraphrase matching**: Multi-word expressions 4. **Recall**: Computes F-score (precision + recall) 5. **Chunk penalty**: Penalizes fragmented matches

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty})$$

Where: - $F_{\text{mean}} = \frac{10PR}{R+9P}$ (recall-weighted harmonic mean) - Penalty based on number of chunks

**Alignment:** Creates explicit word-to-word alignment between candidate and reference.

### 3.2 Strengths

1. **Better correlation**: Higher correlation with human judgments (0.7-0.8)
2. **Linguistic awareness**: Handles morphology and synonymy
3. **Recall inclusion**: Penalizes missing content
4. **Tunable**: Parameters can be tuned per language
5. **Sentence-level**: Better sentence-level correlation than BLEU

### 3.3 Weaknesses

1. **Language-dependence**: Requires WordNet/stemmers
2. **Slower**: More computationally expensive than BLEU

3. **Complexity**: Many parameters, less interpretable
4. **Resource requirements**: Needs linguistic databases
5. **Less ubiquitous**: Not always reported in papers

### 3.4 Adoption

Widely used in shared tasks (WMT) as a complementary metric to BLEU. Particularly effective for morphologically rich languages.

---

## 4. TER: Translation Error Rate

### 4.1 Methodology

TER [@snover2006study] measures the minimum number of edits (insertions, deletions, substitutions, shifts) needed to transform candidate into reference:

$$\text{TER} = \frac{\text{Number of edits}}{\text{Average reference length}}$$

**Shift operation:** Allows moving contiguous word sequences (addresses reordering).

**Lower is better** (unlike BLEU where higher is better).

### 4.2 Strengths

1. **Error-centric**: Direct measure of post-editing effort
2. **Interpretable**: Number of edits has clear meaning
3. **Reordering**: Shift operation handles word order changes
4. **Simplicity**: Straightforward edit distance

### 4.3 Weaknesses

1. **Precision-only**: Like BLEU, doesn't measure recall well
2. **Reference-dependent**: Very sensitive to reference wording
3. **Equal edit costs**: All edits weighted equally (unrealistic)
4. **Lower correlation**: Generally lower than METEOR (0.55-0.65)

### 4.4 Variants

- **HTER**: Human-targeted TER (human post-edits as reference)
- **CharacTER**: Character-level TER [@wang2016character]

---

## 5. chrF: Character n-gram F-score

### 5.1 Methodology

chrF [@popovic2015chrf] operates on character level rather than word level:

$$\text{chrF} = \frac{(1 + \beta^2) \cdot \text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}}$$

Where: - chrP = character n-gram precision - chrR = character n-gram recall - $\beta$ = weight parameter (typically $\beta = 2$, emphasizing recall)

**Typical:** Uses character 6-grams

### 5.2 Strengths

1. **Morphology-friendly**: Handles morphological variants naturally
2. **Language-independent**: No tokenization needed
3. **Agglutinative languages**: Excellent for Turkish, Finnish, etc.
4. **Compounding**: Handles German compounds well
5. **Typo-tolerant**: Partial credit for minor spelling differences
6. **No resources**: Requires no linguistic tools

### 5.3 Weaknesses

1. **Word meaning**: Doesn't capture semantics
2. **Optimal n-gram order**: Requires tuning per language
3. **Speed**: Slower than word-level metrics (longer sequences)
4. **Interpretability**: Less intuitive than word-level

### 5.4 Performance

chrF++ [@popovic2017chrf] (chrF + word n-grams) shows correlation comparable to METEOR, especially for morphologically rich languages.

---

## 6. BEER: Better Evaluation as Ranking

### 6.1 Methodology

BEER [@stanojevic2014beer] combines multiple features using a trained model: - Character n-grams - Word n-grams
- Reordering (permutation trees) - Function word weights - **Word embeddings**: Semantic similarity

**Learning-based:** Trained on human judgments using linear regression.

### 6.2 Strengths

1. **Feature combination**: Leverages multiple signals
2. **Embeddings**: Captures semantic similarity
3. **Reordering**: Sophisticated permutation trees
4. **Tunable**: Can optimize for specific language pairs
5. **High correlation**: 0.75-0.80 with human judgments

### 6.3 Weaknesses

1. **Training requirement**: Needs human judgment data
2. **Embedding quality**: Depends on quality of word vectors
3. **Complexity**: Black-box model (less interpretable)
4. **Speed**: Slower than simple metrics
5. **Limited adoption**: Less widely used than BLEU/METEOR

---

## 7. COMET and BLEURT: Neural Evaluation Metrics

### 7.1 Paradigm Shift

Recent metrics use **pre-trained language models** (BERT, XLM-R) to score translations [@rei2020comet; @sellam2020bleurt].

**Key Idea:** Fine-tune multilingual encoders on human judgment data.

### 7.2 COMET (Crosslingual Optimized Metric for Evaluation of Translation)

**Architecture:** 1. Encode source, candidate, and reference with XLM-RoBERTa 2. Concatenate representations 3. Feed-forward layers predict human score

**Training:** Direct Assessment scores from WMT shared tasks

**Variants:** - **COMET-MQM**: Trained on MQM (Multidimensional Quality Metrics) - **COMET-QE**: Quality estimation (no reference needed)

### 7.3 BLEURT (Bilingual Evaluation Understudy with Representations from Transformers)

**Approach:** 1. Pre-train BERT on synthetic data (backtranslation, masking) 2. Fine-tune on human ratings (WMT) 3. Predict translation quality score

**Pre-training tasks:** - Masked token recovery - Backtranslation ranking - Entailment

### 7.4 Strengths of Neural Metrics

1. **State-of-the-art correlation**: 0.85-0.90 with humans (segment-level)

2. **Semantic understanding**: Captures meaning, not just surface form
3. **Multilingual**: Single model for many language pairs
4. **Context-aware**: Understands discourse and context
5. **Paraphrase-robust**: Recognizes semantic equivalence

### 7.5 Weaknesses of Neural Metrics

1. **Computational cost**: Requires GPU, slow (100-1000× slower than BLEU)
2. **Black-box**: Difficult to interpret or debug
3. **Data dependency**: Needs large human judgment datasets
4. **Reproducibility**: Model versions, training data affect scores
5. **Brittleness**: Can be fooled by adversarial examples
6. **Resource requirements**: Large models (GB of parameters)

### 7.6 Adoption

Increasingly used in research (WMT shared tasks use COMET as primary metric since 2020), but **BLEU still dominant in production** due to speed and simplicity.

---

## 8. Comparative Analysis

### 8.1 Correlation with Human Judgments

| Metric | Segment-Level | System-Level | Resource Needs | Speed |
|---|---|---|---|---|
| **BLEU** | 0.35-0.45 | 0.65-0.75 | None | Very Fast |
| **METEOR** | 0.50-0.60 | 0.70-0.80 | Medium | Fast |
| **TER** | 0.40-0.50 | 0.60-0.70 | None | Fast |
| **chrF** | 0.45-0.55 | 0.65-0.75 | None | Medium |
| **BEER** | 0.55-0.65 | 0.75-0.80 | High | Medium |
| **COMET** | 0.75-0.85 | 0.85-0.90 | Very High | Very Slow |
| **BLEURT** | 0.70-0.80 | 0.80-0.88 | Very High | Very Slow |

*Values are approximate Pearson's r from WMT meta-evaluations [@mathur2020tangled]*

### 8.2 Metric Comparison Table

| Metric | Core Idea | Key Strength | Key Weakness | Correlation |
|---|---|---|---|---|
| **BLEU** | N-gram precision + BP | Simple, fast, ubiquitous | Ignores semantics, low segment-level | Medium |

| Metric | Core Idea | Key Strength | Key Weakness | Correlation |
|---|---|---|---|---|
| **METEOR** | Alignment with linguistic resources | Handles synonyms/stems | Language-dependent | Medium-High |
| **TER** | Edit distance | Intuitive (edit count) | Precision-biased | Medium |
| **chrF** | Character n-grams | Morphology-robust | Less interpretable | Medium |
| **BEER** | Learned combination | Embeddings + features | Needs training data | High |
| **COMET** | Neural (XLM-R) | SOTA correlation | Slow, black-box | Very High |
| **BLEURT** | Neural (BERT) | Semantic understanding | Resource-heavy | Very High |

### 8.3 When to Use Each Metric

**BLEU:** - Rapid prototyping - System comparison (corpus-level) - Historical comparison - Resource-constrained environments

**METEOR:** - Better sentence-level evaluation - Morphologically rich languages - When linguistic resources available

**TER:** - Post-editing effort estimation - Error analysis - Commercial translation

**chrF:** - Low-resource languages - Morphologically complex languages - No tokenization available

**BEER/COMET/BLEURT:** - Research (WMT shared tasks) - High-stakes evaluation - When computational resources available - Segment-level quality estimation

---

## 9. Why BLEU Persists Despite Limitations

Despite superior alternatives, BLEU remains dominant for several reasons:

### 9.1 Historical Inertia

**Published in 2002**, BLEU was the first widely-adopted automatic metric. 20+ years of papers report BLEU, making it the baseline for comparison.

**Community consensus:** Everyone reports BLEU, so everyone continues to report BLEU (network effect).

### 9.2 Simplicity and Speed

- **10 lines of code** to implement basic BLEU
- **Milliseconds** to compute (vs. minutes for neural metrics)
- **No dependencies**: No linguistic resources, no GPU

**Practical impact:** Enables rapid experimentation and debugging.

### 9.3 Interpretability

BLEU's components are intuitive: - N-gram precision: "How many words/phrases match?" - Brevity penalty: "Is the translation complete?"

Neural metrics produce opaque scores (What does "COMET: 0.73" mean?).

### 9.4 Reproducibility

BLEU is deterministic and version-stable. `sacreBLEU` [@post2018call] standardized implementation, ensuring reproducibility.

Neural metrics have: - Model version differences - Training data differences - Stochastic variation (dropout)

### 9.5 Corpus-Level Reliability

While BLEU is poor at segment-level, it's **reliable at corpus-level** (0.65-0.75 correlation), which is sufficient for system comparison [@post2018call].

### 9.6 Limitations Are Well-Understood

Researchers know BLEU's weaknesses and account for them: - Report multiple metrics (BLEU + METEOR + TER) - Use human evaluation for final results - Recognize BLEU as a "sanity check"

### 9.7 Resistance to Change

Academia and industry have substantial infrastructure built around BLEU: - Evaluation scripts - Benchmarks with BLEU scores - Expectations ("NMT systems achieve ~35 BLEU on WMT14")

Changing to new metrics requires recalibrating expectations.

---

## 10. Conclusion

### Summary of Findings

Automatic MT evaluation has evolved from simple n-gram overlap (BLEU) to sophisticated neural models (COMET, BLEURT). Each metric represents a trade-

off between **accuracy** (correlation with humans), **speed**, **interpretability**, and **resource requirements**.

**Key Insights:**

1. **No perfect metric exists**: All metrics have weaknesses
2. **Context matters**: Best metric depends on use case (research vs. production)
3. **Multiple metrics recommended**: Ensemble evaluation more robust [@ma2019results]
4. **Human evaluation remains gold standard**: Automatic metrics are proxies
5. **BLEU persists for good reasons**: Despite limitations, its simplicity and ubiquity make it irreplaceable for now

**Future Directions**

1. **Reference-free evaluation**: QE metrics (COMET-QE) that don't need references
2. **Explainable neural metrics**: Interpretability for learned metrics
3. **Task-specific metrics**: Evaluation aligned with downstream tasks
4. **Multilingual metrics**: Single metric for all language pairs
5. **Efficient neural metrics**: Distilled models for faster computation

**Recommendations**

**For practitioners:** - **Report multiple metrics** (BLEU + METEOR/chrF + neural metric if possible) - **Use BLEU for development**, neural metrics for final evaluation - **Include human evaluation** for publication-quality claims - **Use `sacreBLEU`** for reproducible BLEU scores

**For researchers:** - Continue developing better automatic metrics - Focus on interpretability and efficiency - Validate on diverse languages and domains - Release standardized implementations

**Final Thought**

BLEU revolutionized MT evaluation when introduced in 2002. While newer metrics show higher correlation with human judgments, BLEU's simplicity, speed, and widespread adoption ensure it will remain the de facto standard for the foreseeable future. The key is understanding its limitations and using it appropriately as **one tool among many** in the evaluation toolkit.

---

## References (IEEE)

[1] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311-318.

[2] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop*, 2005, pp. 65-72.

[3] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. AMTA*, 2006.

[4] M. Popovic, "chrF: Character n-gram F-score for automatic MT evaluation," in *Proc. WMT*, 2015, pp. 392-395.

[5] T. Sellam, D. Das, and A. P. Parikh, "BLEURT: Learning robust metrics for text generation," *arXiv preprint arXiv:2004.04696*, 2020.

[6] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in *Proc. EMNLP*, 2020, pp. 2685-2702.

[7] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. WMT*, 2014, pp. 376-380.

[8] P. Koehn, *Statistical Machine Translation*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[9] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of BLEU in machine translation research," in *Proc. EACL*, 2006, pp. 249-256.

[10] N. Mathur, T. Baldwin, and T. Cohn, "Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics," in *Proc. ACL*, 2020, pp. 4984-4997.

[11] O. Bojar et al., "Results of the WMT16 metrics shared task," in *Proc. WMT*, 2016, pp. 199-231.

[12] M. Post, "A call for clarity in reporting BLEU scores," in *Proc. WMT*, 2018, pp. 186-191.

[13] M. Freitag, D. Grangier, and I. Caswell, "BLEU might be guilty but references are not innocent," in *Proc. EMNLP*, 2020, pp. 61-71.

[14] M. Popovic, "chrF++: Words helping character n-grams," in *Proc. WMT*, 2017, pp. 612-618.