Referenceless evaluation of machine translation models by ranking performance in Romanian to English translate-train settings

Mihail Feraru[1], **Alexandra Diaconu**[1], **Bogdan Alexe**[1,2]
[1]University of Bucharest
Str. Academiei 14, 010014 Bucharest, Romania
[2]Gheorghe Mihoc-Caius Iacob Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy
Calea 13 Septembrie 13, 050711 Bucharest, Romania
mihailferaru2000@gmail.com, {alexandra.diaconu, bogdan.alexe}@fmi.unibuc.ro

Abstract

We propose a referenceless evaluation method for machine translation (MT) models by assessing their performance in *translate-train* scenarios across a variety of natural language processing (NLP) tasks. The approach ranks MT systems based on the downstream impact of their translations on independent NLP models trained on translated data, thus eliminating the need for professional ground-truth references. We evaluate four prominent MT tools — ChatGPT 3.5 Turbo, DeepL, Google Translate, and Mistral 7B Instruct v0.2 — on the Romanian→English language pair and analyze their influence on *text summarization*, *sentiment analysis*, and *authorship identification*. To further test the generalization and robustness of our method, we extend the evaluation to a cross-modality setup using *out-of-domain speech data*. In this setting, speech segments are transcribed with Whisper-Large, translated into English, and used in a four-class domain classification task (children's stories, audiobooks, film dialogues, podcasts). Our findings show that translation improves downstream performance for sentiment analysis and summarization, while stylistically rich texts such as poetry or noisy ASR transcriptions suffer degradation. The proposed ranking metric correlates strongly with human judgments and remains sensitive to translation quality even in multimodal pipelines, providing a scalable and practical alternative to reference-based MT evaluation.

Keywords: machine translation evaluation, translate-train, referenceless metrics

1. Introduction

Reliable evaluation of machine translation (MT) systems remains an open challenge despite extensive prior work. Existing metrics often depend on the availability and quality of reference translations, which are costly to obtain and can introduce bias due to translator style, reference noise, or domain mismatch. Furthermore, such reference-based metrics (e.g., BLEU or chrF) may not accurately capture how translation quality affects practical downstream tasks.

Recent research has increasingly emphasized evaluation methods that do not rely on human-crafted references, seeking to measure translation quality from alternative perspectives such as semantic consistency, fluency, or downstream task performance (Lee et al., 2023). The main use case of evaluation metrics is to provide a reliable comparison between MT systems, motivating approaches that can operate without parallel corpora.

We propose a *referenceless evaluation framework* for MT systems based on the *translate-train* paradigm. Instead of comparing translated outputs to human references, we evaluate how translations affect downstream NLP task performance after the source-language datasets are translated into the target language. In essence, translation quality is inferred from the impact it has on independent models fine-tuned on the translated data.

We apply our framework on the *Romanian→English* language pair, a relatively low-resource direction that has received little attention in translate-train settings. Using three textual NLP tasks, *sentiment analysis*, *text summarization*, and *authorship identification in poetry*, we quantify how different MT systems influence model performance when trained on translated data.

To further test the *robustness and generalization* of the proposed method beyond text-only inputs, we extend our experiments to *out-of-domain speech data* from the RO-N3WS corpus (Diaconu et al., 2026). In this setting, speech segments are automatically transcribed in Romanian using the *Whisper-Large* model (Radford et al., 2023), translated into English by the same MT systems, and used in a four-class domain classification task covering children's stories, audiobooks, film dialogues, and podcasts. This cross-modality extension allows us to examine whether our evaluation metric remains sensitive to translation quality even when upstream inputs are generated by automatic speech recognition.

The main contributions of this paper are as follows: (1) We propose a *referenceless evaluation method* for MT systems that ranks models based on their downstream impact in translate-train sce-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8813–8823
11–16 May 2026.

narios, without requiring human reference translations; (2) We conduct the first comprehensive *Romanian→English translate-train study*, covering sentiment analysis, summarization, and authorship detection tasks, and analyze which settings benefit from translation and which degrade in performance; (3) We demonstrate that our ranking metric *correlates strongly with human judgments* while maintaining robustness across modalities through an additional cross-modality experiment on RO-N3WS speech data.

2. Background

This section summarizes key concepts underlying our referenceless evaluation approach and the translate-train paradigm.

Human evaluation of machine translation encompasses a broad range of methodologies for assessing translation quality in a systematic manner. As highlighted by Freitag et al. (2021), these include rating translation quality on discrete or continuous scales at the segment or document level, as well as identifying or annotating specific errors (e.g., lexical, syntactic, or stylistic). Some methods also assess comprehension via gap-filling or reading tasks. However, scale-based judgments often suffer from high variability due to annotator subjectivity and inconsistency across raters.

Reference-based MT evaluation relies on the availability of one or more human-crafted reference translations, typically produced by professional translators, to compute similarity-based metrics such as BLEU, chrF, or COMET (Lee et al., 2023). While these metrics remain the dominant standard, they are inherently limited by the quality, style, and representativeness of the reference text, as well as by the availability of parallel corpora for low-resource language pairs.

Quality Estimation (QE) emerged as a reference-free alternative, initially proposed in the WMT19 shared task (Fonseca et al., 2019). QE systems estimate the quality of an MT output directly, without access to a reference translation, typically by leveraging supervised or semi-supervised models trained to predict sentence-level quality scores. These approaches inspired later referenceless evaluation techniques that aim to capture translation adequacy and fluency intrinsically, rather than through comparison to a gold standard.

Translate-train is a cross-lingual learning strategy designed to mitigate data scarcity in low-resource languages. The idea, explored by Jundi and Lapesa (2022) and Artetxe et al. (2023), is to translate available datasets from a low-resource source language into a high-resource target language and then fine-tune task-specific models (e.g., for classification or summarization) on the translated data. This technique enables leveraging strong pre-trained target-language models while preserving task semantics. Further extensions, such as Yang et al. (2024), apply translate-train in knowledge distillation settings to train bilingual or multilingual encoders from monolingual ones.

Overall, translate-train provides a natural foundation for our work: by observing how translation affects the downstream performance of NLP models trained on translated data, we obtain an indirect yet informative signal about translation quality, without relying on human-generated references.

3. Related Work

Limitations of reference-based metrics. A large body of research has analyzed the shortcomings of traditional automatic metrics that rely on human-crafted reference translations. Results from the WMT22 shared task (Freitag et al., 2020) show that such metrics often exhibit low correlation with human judgments across diverse language pairs. Classical match-based metrics like BLEU, chrF, or TER tend to underestimate the quality of high-performing MT systems, especially when multiple valid translations exist (Mathur et al., 2020; Reiter, 2018). Even small variations in lexical choice or phrasing can yield large score differences. As highlighted by Kocmi et al. (2021), this reliance on surface-level similarity has limited the progress of MT evaluation for many years.

Impact of reference quality. Recent studies have also demonstrated that the quality and consistency of reference translations strongly influence evaluation outcomes. Zouhar and Bojar (2024) investigated groups of translators with varying expertise and found that the highest correlation with human judgment was obtained not for the most expert translators, but for an intermediate-quality group. This surprising finding underscores the instability of reference-based metrics and the inherent bias introduced by reference selection and quantity.

Referenceless and quality estimation approaches. To overcome these limitations, recent work has explored reference-free evaluation methods that directly model translation quality. The WMT19 shared task on Quality Estimation (QE) (Fonseca et al., 2019) established a foundation for predicting sentence-level MT quality without references. Building on this line of work, several methods exploit multilingual language models to capture cross-lingual semantics.

YiSi-2 with bilingual mappings (Lo and Larkin, 2020) computes semantic similarity between source and target sentences using multilingual contextual embeddings (e.g., BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020)). By aligning embeddings across languages via bilin-

gual projection, YiSi-2 improves correlation with human ratings, addressing the language-clustering effect that weakens cross-lingual comparisons.

Target-side language modeling (Zhang et al., 2022) offers a purely monolingual alternative: it evaluates MT output fluency by measuring perplexity under a pretrained target-language model such as XLM-R. This approach simplifies evaluation but captures only fluency, not adequacy.

Cross-lingual embedding alignment, proposed by Zhang et al. (2023), uses multilingual knowledge distillation to implicitly align representations of parallel sentences, later integrated into metrics such as BERTScore and Word Mover's Distance. Such methods enhance semantic equivalence estimation without explicit references.

COMET and COMET-QE (Rei et al., 2020) represent a major leap toward neural MT evaluation. COMET-QE is trained on human-annotated QE datasets and predicts segment-level quality scores using representations of both source and translation. It achieves a strong correlation with human judgments and has become a standard baseline for referenceless evaluation.

Our position. While prior reference-free methods rely on sentence-level semantic similarity or quality estimation models, our approach is fundamentally different: we assess translation quality *extrinsically*, by quantifying how translated data influence downstream task performance in a translate-train setting. This perspective shifts the focus from intrinsic text similarity to the *functional utility* of translations in practical NLP pipelines. In contrast to previous metrics, our framework can be applied with generic monolingual datasets in the source language and extends naturally to multimodal settings, as demonstrated by our cross-modality experiments on RO-N3WS speech data.

4. Proposed referenceless evaluation

We propose a reference-free evaluation framework that compares machine translation (MT) systems by measuring their impact on the performance of downstream transformer-based models in supervised NLP tasks. Rather than assessing the similarity between translations and human-crafted references, our method quantifies how much the use of translated data affects task performance under the *translate-train* paradigm. This approach eliminates variability caused by reference quality and reduces the cost of data acquisition while maintaining high correlation with human judgments.

Problem setup. Let us consider a fixed language pair (source $\rightarrow$ target) and a set of n translation systems $T_j,\ j \in \overline{1,n}$ to be compared. Instead of relying on parallel corpora, we use m *monolingual datasets* in the source language, denoted by $\mathcal{D}_i^s,\ i \in \overline{1,m}$, each associated with a supervised NLP task such as classification or summarization. Each dataset has a bounded evaluation function $f_i(y, \hat{y})$ (e.g., F1-score or ROUGE-L), which we denote more compactly as $f_{\mathcal{D}_i^s}(\mathcal{M})$ for a model $\mathcal{M}$. For every dataset $\mathcal{D}_i^s$ and translator T_j, we generate a translated version in the target language:

$$\mathcal{D}_{ij}^t = T_j(\mathcal{D}_i^s).$$

We then train two task-specific transformer models with identical architectures: $\mathcal{M}_i^s$ (for the source language) and $\mathcal{M}_{ij}^t$ (for the translated dataset). This design ensures that any performance difference can be attributed to translation quality rather than model capacity.

Evaluation procedure. For each dataset $\mathcal{D}_i^s$, the following steps are performed:

1. Split $\mathcal{D}_i^s$ into train and evaluation subsets (or apply k-fold cross-validation);

2. Train and evaluate $\mathcal{M}_i^s$ on $\mathcal{D}_i^s$ to obtain a source-language baseline score $f_i(\mathcal{M}_i^s)$;

3. Train and evaluate $\mathcal{M}_{ij}^t$ on $\mathcal{D}_{ij}^t$ following the same protocol to obtain the translated-data score $f_i(\mathcal{M}_{ij}^t)$;

4. Compute the *performance difference*:

$$\Delta(\mathcal{D}_i^s, T_j) = f_i(\mathcal{M}_i^s) - f_i(\mathcal{M}_{ij}^t), \qquad (1)$$

   which captures the degradation (or improvement) in task performance induced by translation. A smaller Δ indicates better translation quality, as the translated dataset preserves task-relevant semantics more faithfully;

5. Aggregate the differences across all datasets to compute a global score for each translator:

$$S(T_j) = \sum_{i=1}^{m} \Delta(\mathcal{D}_i^s, T_j). \qquad (2)$$

Intuition. The metric $\Delta(\mathcal{D}_i^s, T_j)$ reflects how translation affects model learning for a given task, while the cumulative score $S(T_j)$ provides a robust system-level ranking across diverse datasets. This formulation enables fair comparison among MT systems even when no human reference translations exist, and it naturally extends to multimodal or noisy inputs, as demonstrated later with our RO-N3WS cross-modality experiments.

5. Translate-train on Romanian to English

This section details the experimental setup used to implement the proposed referenceless evaluation framework on the Romanian→English language pair. We describe the selected datasets, the associated NLP tasks and metrics, the translation and transformer models employed, and the overall training and evaluation setup. All experiments were conducted under the same translate-train protocol to ensure comparability across MT systems.

5.1. Datasets

Selection criteria. We surveyed available Romanian NLP datasets covering various tasks such as sentiment analysis, summarization, fake news detection, authorship attribution, and others. Two main criteria guided the selection: (1) the task must remain well defined after translation to English, and (2) the dataset must contain at least ten thousand samples to enable robust fine-tuning. Tasks such as named entity recognition were excluded because label transfer across languages is non-trivial. Based on these criteria, we selected three datasets described below.

RoSent (Dumitrescu et al., 2020) consists of 28,000 movie and product reviews annotated with positive or negative sentiment labels. A stratified random subsample of 4,000 instances was used for our experiments.

RoTextSummarization (Niculescu et al., 2022) contains approximately 72,000 Romanian news articles and their human-written summaries collected between 2020 and 2022. A genre-balanced subset of 8,000 examples was used.

Rupert[1] is a corpus of Romanian poetry containing more than 17,000 poems by over 500 authors. To ensure sufficient class representation, we retained the 25 most frequent authors and sampled 5,000 poems stratified by author identity.

Subsampling. For each dataset, we used approximately 10–30% of the total data (stratified where applicable) to balance computational cost and statistical stability. Translation costs were also a consideration, as private APIs such as DeepL and ChatGPT incur usage fees per character.

Qualitative Examples. To better illustrate the differences between translation systems, we present qualitative examples from the two datasets used in our experiments. Table 1 shows an example from the *RoSent* sentiment analysis dataset, while Table 2 provides a parallel comparison of translations for a poetic excerpt from the *Rupert* authorship identification dataset.

In the RoSent example (Table 1), the Romanian source text contains informal language, minor grammatical inconsistencies, and colloquial phrasing typical of user-generated movie reviews. All MT systems capture the overall sentiment of the text, but they differ in lexical choices and fluency. The large language model systems (ChatGPT and Mistral) tend to produce slightly more fluent sentences, occasionally normalizing punctuation and capitalization. In contrast, the classical MT systems (Google Translate and DeepL) remain closer to the structure of the source text but sometimes propagate lexical errors present in the original sentence.

Table 2 highlights a more challenging case involving poetic text from Mihai Eminescu's poem *Ce e amorul?*. Unlike the RoSent example, poetic translation introduces additional difficulties due to the presence of rhyme, meter, and stylistic devices. The parallel layout of the table allows a line-by-line comparison of the Romanian verses and their English translations. While all systems broadly preserve the semantic content of the poem, none of them maintain the original rhyme or metrical structure. Furthermore, some lexical and grammatical inconsistencies appear across translations, such as incorrect pronoun usage (e.g., translating *amorul* with gendered pronouns) or subtle mistranslations of context-dependent words.

Overall, these examples illustrate how translation quality interacts with downstream tasks. In sentiment analysis, preserving general semantic polarity is sufficient for good performance, whereas in authorship identification the loss of stylistic features such as rhythm, lexical nuance, and syntactic structure can significantly affect model performance.

5.2. Tasks and metrics

Sentiment analysis (RoSent) involves classifying reviews as positive or negative. We report the *macro-averaged F1-score*, which equals accuracy for balanced data but is more robust to imbalances.

Text summarization (RoTextSummarization) is a sequence-to-sequence task aiming to generate a summary of a given article. Performance is measured using *ROUGE-L*, the most widely used overlap-based metric for summarization.

Authorship identification (Rupert) is a multiclass classification task where the model predicts the author of a given poem. This task emphasizes stylistic and lexical cues that are often lost during translation. We again use the *macro-averaged F1-score* as the evaluation metric.

[1] https://huggingface.co/datasets/littlewho/Rupert

Model	Romanian → English Example (RoSent)
Original	chucky e inapoi ... si este timpul! de aceasta data, cu ajutorul lui jennifer tilly si o vraja putin de la voodoo for dummies. ei bine, cel putin cu aceasta transa, tabara se intoarce [...]
ChatGPT 3.5	chucky is back... and it's time! this time, with the help of jennifer tilly and a little voodoo spell from *voodoo for dummies*. this has been by far the most fun in the series [...]
Mistral 7B	chucky is back ... and it's about time! with the help of jennifer tilly and a little voodoo from *voodoo for dummies*. without a doubt the most entertaining installment [...]
Google Translate	chucky is back... and it's about time! this time, with the help of jennifer tilly and a little spell from voodoo for dummies [...]
DeepL	chucky's back... and it's time! this time, with the help of jennifer tilly and a little spell from voodoo for dummies [...]

Table 1: Example from the RoSent dataset showing Romanian source text and English translations produced by the evaluated MT systems.

Romanian	ChatGPT 3.5	Mistral 7B	Google Translate	DeepL
Ce e amorul? E un lung	What is love? It's a long	What is love? It is a long	What is love? It's a long	What is love? It's a long
Prilej pentru durere,	opportunity for pain,	privilege for pain,	time for pain,	time for pain,
Căci mii de lacrimi nu-i ajung	Because thousands of tears	For thousands of tears	Because thousands of tears	For a thousand tears
Şi tot mai multe cere.	are not enough and it asks	are not enough and it asks	are not enough and he asks	are not enough and more asks
De-un semn în treacăt de la ea	With a fleeting sign from her	From a passing sign from her	With a passing sign from her	From a passing sign from her
El sufletul ţi-l leagă,	it binds your soul	it binds your soul	he binds your soul	he binds your soul
Încât să n-o mai poţi uita	So that you cannot forget her	So that you cannot forget	So that you can no longer forget	So that thou canst not forget
Viaţa ta întreagă.	your whole life.	your entire life.	your whole life.	her your whole life.
Dar încă de te-aşteaptă-n prag	But even if she waits	Yet it waits for you	But still waiting for you	But still she waits
În umbră de unghere,	in the shadow of corners	in the shadow of corners	in the shadow of the corner	in the shadows of corners
De se-ntâlneşte drag cu drag	If love meets love	When love meets love	To meet love with love	If love meets love
Cum inima ta cere:	as your heart desires	as your heart desires	as your heart asks	as your heart asks
Dispar şi ceruri şi pământ	Heaven and earth disappear	The sky and earth disappear	Heaven and earth disappear	Heaven and earth disappear
Şi pieptul tău se bate,	and your chest beats	and your chest beats	and your chest beats	and thy breast beats
Şi totu-atârnă de-un cuvânt	And everything depends on	And you are suspended from	And everything hangs on	And all hangs on
Şoptit pe jumătate.	a word whispered halfway.	a word whispered half-heartedly.	a half-whispered word.	a word half whispered.

Table 2: Parallel comparison of a poetic excerpt from the Rupert dataset (Mihai Eminescu, *Ce e amorul?*) and its translations produced by the evaluated MT systems. Lines are aligned across columns to facilitate comparison of lexical and stylistic differences.

5.3. Machine translation models

We evaluated four translation systems representing both commercial MT services and LLMs.

Google Translate and **DeepL** are industry-standard systems optimized for high-quality translations across many language pairs. Both were accessed via their public APIs and used in sentence-level mode (no batching).

ChatGPT 3.5 Turbo[2] and **Mistral 7B Instruct v0.2** (Jiang et al., 2023) are general-purpose large language models capable of translation. ChatGPT was prompted in a zero-shot manner using: *"Translate from Romanian to English: <source text>"*, while Mistral 7B required a lightweight prompt-engineering step to prevent hallucinations: *You*

[2] https://openai.com/index/chatgpt/

are a helpful professional translator. You will be prompted with texts to translate. You will respond only with the translation. You will receive prompts with the format: "Translate from Romanian to English: [Romanian text]". You will respond with: "Translation: [English text]. Translate from Romanian to English: <source text>. Running Mistral in 16-bit quantized mode yielded only 0.5% non-conforming outputs. We ignored rare hallucinations that still produced syntactically valid outputs, as they did not affect downstream evaluation.

5.4. Transformer models

For downstream NLP tasks, we selected pretrained transformer architectures with comparable capacity across source and target languages.

BERT (Devlin et al., 2019) was used for classification tasks (sentiment and authorship). We employed the English *BERT Base Cased* model (110M parameters) pre-trained on a 60 GiB corpus, and its Romanian counterpart fine-tuned by Dumitrescu et al. (2020) on a 15 GiB corpus.

BART (Lewis et al., 2020) was used for summarization, leveraging the English base model and a Romanian variant trained from scratch on a 50,GiB corpus.[3]Each model has about 140M parameters and a maximum output length of 1024 tokens.

5.5. Training setup

For all datasets, we trained models with 5-fold cross-validation using consistent folds across Romanian and translated versions. Texts were tokenized and truncated to the 95th percentile of sequence length, reducing training time by about 50%. We fine-tuned each model for 10 epochs with unfrozen layers using the Hugging Face framework and AdamW optimizer (5×10^{-5} learning rate, $\beta_1 = 0.9$, $\beta_2 = 0.999$, no weight decay). Batch sizes were adapted per dataset (RoTextSummarization: 8, Rupert: 16, RoSent: 32) to fully utilize GPU memory. All experiments ran on NVIDIA RTX 4090 GPUs, totaling about 150 compute hours.

5.6. Results

Table 3 reports average task performance across five folds for each MT model and dataset. The *baseline diff* column corresponds to $\Delta(\mathcal{D}_i^s, T_j)$ in Eq. 1, and the final column aggregates these differences into the overall system score $S(T_j)$ (Eq. 2).

Performance boost on basic tasks. The translate-train approach improved downstream performance for sentiment analysis (in three of four MT systems) and for all systems in text summarization. We note that ROUGE-L may overestimate performance due to overlap with translated references, but the pattern aligns with prior cross-lingual studies (Jundi and Lapesa, 2022; Artetxe et al., 2023). These results suggest that Romanian NLP tasks with primarily lexical content (e.g., sentiment) benefit from translation into English.

Performance degradation on stylistic tasks. For authorship identification in poetry, all MT systems performed worse than the Romanian baseline, reflecting the loss of stylistic and rhythmic cues during translation. This confirms that tasks requiring deeper semantic understanding are less suited for translate-train evaluation.

Cross-task variability. The best-performing translator varies across tasks, indicating that different MT systems exhibit complementary strengths. Such variation underscores the importance of a task-aggregated score like $S(T_j)$ for fair ranking. We later extend this framework beyond textual data by applying the same methodology to out-of-domain speech transcriptions from the RO-N3WS corpus (Section 7), demonstrating that our metric generalizes to multimodal inputs.

6. Results of MT models evaluation

Last column of Table 3 presents the overall scores assigned to each MT system according to Equation 2. Each score represents the cumulative sum of performance differences across all NLP tasks between the translate-train setup and the corresponding Romanian baseline.

6.1. Human Judgement Collection

To validate the proposed referenceless metric against human perception, we conducted a human evaluation based on pairwise translation preferences. A subset of 240 Romanian texts was randomly sampled from the three NLP datasets (RoSent, RoTextSummarization, and Rupert). Each text was translated by two of the four MT systems and presented side by side to annotators, who indicated their preferred translation or selected a tie using a five-point slider ranging from "left much better" to "right much better." MT system identities were hidden to prevent bias, and both text order and translation order were randomized.

Each annotator received a unique questionnaire of 60 text pairs, balanced across the three datasets. In total, we collected about 900 individual judgments from 15 volunteers, including students and professors in computer science, law, and foreign languages. All participants were native Romanian

[3] https://huggingface.co/Iulian277/ro-bart-1024

	RoSent		RoTextSummarization		Rupert		
	f1-macro	baseline diff	rouge-l	baseline diff	f1-macro	baseline diff	**our score**
Translate-Train with English results							
ChatGPT3.5	92.33±00.69	+01.08	**30.33±00.60**	**+06.07**	**65.28±01.71**	**-05.78**	+01.37
GTranslate	91.82±00.49	+00.57	30.32±00.27	+06.06	64.88±01.98	-06.18	+00.45
DeepL	**92.35±00.98**	**+01.11**	29.23±00.51	+04.97	63.66±03.00	-07.40	-01.32
Mistral 7B	90.34±00.49	-00.91	28.08±00.64	+03.82	58.17±01.28	-12.89	-09.98
Romanian results							
Romanian	91.24±00.30	-	24.26±00.43	-	71.06±01.19	-	

Table 3: Training and evaluation results of baseline and translate-train experiments for each dataset in combination with each Machine Translation model. Scores are in range [0, 100] and represent the average over a 5-fold cross-validation run for each result having their standard deviations also reported.

speakers with verified English proficiency at B2–C1 level (reading C1 or higher).

To aggregate votes, we used a simple scoring scheme: each tie contributed 0.5 points to both systems, while a preference added 1 point to the favored system. We also tested an extended 0.5/1.0/1.5 weighting scheme to reflect preference intensity, but results showed no significant difference, so we retained the simpler scoring rule. Table 4 reports the aggregated results alongside our automatic evaluation for comparison.

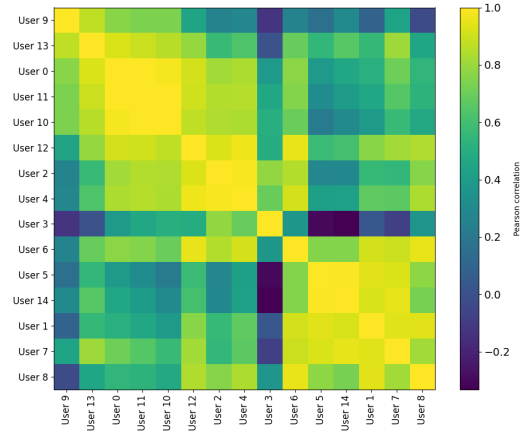

Figure 1: Matrix of pairwise Pearson correlation coefficients between annotators' votes. Clusters indicate groups with higher agreement.

6.2. Inter-Annotator Consistency

To assess consistency among annotators, we computed pairwise Pearson correlation coefficients between individual vote vectors. Figure 1 shows the resulting correlation matrix, reordered to highlight annotator clusters. We observed generally positive correlations, with one outlier (User 3) showing weaker agreement. Descriptive statistics yielded a mean correlation of 0.62, median 0.69, and standard deviation 0.30, indicating moderate to strong overall consistency. The high variance is expected for subjective translation judgements.

6.3. Correlation Between Human and Automatic Evaluation

We next compared the system rankings derived from our referenceless evaluation with those obtained from human preferences. Table 5 reports Pearson correlations at both dataset and system levels. The overall correlation of $r = 0.87$ demonstrates strong alignment between our metric and human judgments, confirming the reliability of the proposed approach. Interestingly, the weakest correlation was observed for the simpler RoSent dataset ($r = 0.48$), suggesting that datasets with limited linguistic complexity may be less informative for distinguishing high-quality MT systems. Conversely, the poetry-based Rupert dataset showed the strongest correlation ($r = 0.98$), indicating that semantically and stylistically rich texts offer a more sensitive signal for evaluating translation quality.

7. Cross-Modality Robustness Evaluation

To further test the robustness and generalization of our referenceless evaluation framework beyond text-only settings, we extended experiments to spoken data using the *RO-N3WS* corpus (Diaconu et al., 2026). This experiment examines whether the same translate-train formulation can capture translation-induced performance differences when the source text originates from automatic speech recognition (ASR) rather than written input.

RO-N3WS corpus. RO-N3WS is a benchmark Romanian speech dataset totaling over 126 hours of manually transcribed audio, designed to study robustness and domain generalization in low-resource ASR. It consists of two major parts: (i) an *in-domain* broadcast news component of approximately 105 hours, collected from national television networks (*ProTV* and *Antena 1*), and (ii) an *out-of-distribution (OOD)* component of about 21 hours, spanning four stylistically distinct domains: audiobooks, film dialogues, children's stories, and podcasts. All segments were manually cleaned and

Translator	RoSent	RoTextSummarization	Rupert	Total Human score	Our score
ChatGPT3.5	81.5	135	73.5	288	+01.37
GTranslate	64	97.5	71.5	227	+00.45
DeepL	57.5	90	54.5	202	-01.32
Mistral 7B	27	89.5	28.5	153	-09.98

Table 4: Aggregated human preference scores for each MT system using a 0.5/1.0 scoring scheme, compared to the automatic scores derived from our proposed evaluation method.

Dataset	Pearson Correlation
RoSent	0.4756
RoTextSummarization	0.6423
Rupert	0.9769
All	0.8741

Table 5: Pearson correlation between human judgments and our referenceless evaluation scores, reported per dataset and at the system level.

annotated following a uniform transcription protocol to ensure consistent orthography, diacritics, and number normalization. The OOD subset contains more than 11,700 speech clips (mean length $\approx$ 6.3 s) covering both acted and spontaneous speech under diverse acoustic conditions, making it particularly suitable for robustness analysis.

Experimental setup. We evaluate the robustness of our referenceless evaluation framework in a cross-modality setting using the OOD portion of the RO-N3WS corpus. The task is formulated as a 4-class *domain classification* problem, where each speech segment must be assigned to one of four domains: children's stories, audiobooks, film dialogues, or podcasts. We consider two types of textual inputs derived from the speech data. First, we use the *ground-truth Romanian transcripts* provided in the dataset. Second, we use *automatic transcripts* produced by the *Whisper-Large* ASR model. In both cases, the resulting Romanian texts are optionally translated into English using the same four MT systems evaluated in the text-only experiments (*ChatGPT 3.5 Turbo*, *DeepL*, *Google Translate*, and *Mistral 7B Instruct v0.2*). For each setting, we fine-tune a *RoBERTa-base* classifier on 4,000 labeled examples using an 80/20 train/test split to predict the speech domain. Performance is measured using the *macro-averaged F1-score*, which provides a balanced evaluation across all classes.

Results and Discussion. Table 6 reports the results for both ground-truth transcripts and Whisper-generated transcripts. Using the original Romanian ground-truth texts yields the highest performance (F1 = 0.856). When these texts are translated into English, performance decreases moderately, with Google Translate achieving the closest result to the Romanian baseline (F1 = 0.801), followed by DeepL, ChatGPT, and Mistral.

When automatic transcripts from Whisper-Large are used instead of ground-truth text, we observe a first degradation due to ASR errors (F1 = 0.856 $\rightarrow$ 0.827). Subsequent translation of these transcripts introduces an additional performance drop, again with Google Translate yielding the best results among the MT systems (F1 = 0.760), followed by DeepL, ChatGPT, and Mistral.

Overall, the relative ranking of MT systems remains consistent across both ground-truth and ASR-derived inputs. This suggests that the proposed referenceless evaluation framework remains sensitive to translation quality even when the source text originates from an automatic speech recognition pipeline. The experiment therefore confirms that our approach generalizes beyond clean text data and can capture performance differences in multimodal speech-to-text processing scenarios.

8. Limitations

Despite the strong correlation with human judgment obtained by our referenceless evaluation framework, several limitations remain, which we aim to address in future work.

Language coverage. All experiments were conducted on a single language pair and direction, *Romanian$\rightarrow$English*. Although this choice highlights the method's suitability for a mid-resource language, it limits generalizability to high-resource or typologically distant pairs. Future work will include additional directions such as English$\rightarrow$Romanian and Romanian$\rightarrow$French to assess cross-lingual transferability.

Scale of human evaluation. The human-judgement study involved a relatively small group of volunteer annotators. While their language proficiency and inter-annotator correlation were satisfactory, the number of raters may not fully represent professional translation standards or large-scale user preferences (Freitag et al., 2021). Expanding the evaluation to a broader pool of participants will improve statistical reliability.

Task coverage and applicability. The translate-train paradigm naturally applies to supervised tasks such as classification or summarization but

Table 6: Cross-modality robustness results on the RO-N3WS OOD speech data. Scores represent mean F1-macro across five runs.

Input Pipeline	Translation Model	F1-macro
Ground-truth Romanian transcripts		
Ground truth (RO text)	—	**0.856**
Ground truth + Google Translate (EN)	GTranslate	0.801
Ground truth + DeepL (EN)	DeepL	0.788
Ground truth + ChatGPT (EN)	ChatGPT 3.5	0.775
Ground truth + Mistral (EN)	Mistral 7B	0.734
Whisper-Large ASR transcripts		
Whisper transcription (RO)	—	**0.827**
Whisper + Google Translate (EN)	GTranslate	0.760
Whisper + DeepL (EN)	DeepL	0.752
Whisper + ChatGPT (EN)	ChatGPT 3.5	0.742
Whisper + Mistral (EN)	Mistral 7B	0.693

is less suitable for sequence-labeling or structured-prediction tasks (e.g., named entity recognition, question answering). This restricts the diversity of datasets available for MT evaluation using our framework. Developing adaptations for partially supervised or generative tasks represents an important direction for extending its applicability.

Cross-modality constraints. The additional experiment on speech-to-text data demonstrates the framework's robustness beyond text, but its scope remains limited to pre-transcribed data. Direct integration with audio-based evaluation or end-to-end speech translation models is left for future exploration. Moreover, errors introduced by automatic speech recognition propagate through the pipeline, potentially conflating ASR and MT quality effects.

Ranking granularity. Our current formulation tends to produce smaller relative differences between high-performing MT systems compared to human assessments. This saturation effect may arise from limited dataset complexity or from the choice of evaluation metrics (e.g., ROUGE-L, F1-macro). Future work will incorporate semantic-level measures such as COMET-QE or BLANC to enhance ranking sensitivity among top-tier systems.

9. Conclusions

We introduced a novel *referenceless evaluation framework* for MT systems, based on measuring their downstream impact in *translate-train* scenarios across multiple NLP tasks. By translating Romanian datasets into English and training task-specific transformer models on the translated data, we quantified translation quality indirectly, through the extent to which translations preserved task-relevant information.

Experiments on three diverse textual tasks, *sentiment analysis*, *text summarization*, and *authorship identification*, revealed that translation improves performance for lexically oriented tasks while degrading performance for stylistically rich ones such as poetry authorship. These results highlight that MT benefits are task-dependent and that translation can amplify or obscure linguistic cues depending on semantic and stylistic complexity.

Our method eliminates the reliance on professional reference translations, thereby reducing evaluation cost and bias while maintaining strong correlation with human judgments ($r = 0.87$ at the system level). The framework consistently ranked MT systems in line with human preferences, confirming its validity as a scalable alternative to traditional metrics such as BLEU or COMET.

To further assess generalization, we extended the framework to a *cross-modality* setting using the RO-N3WS corpus, where speech data were transcribed and translated before downstream classification. The results confirmed that our metric remains sensitive to translation quality even when the source text originates from automatic speech recognition, demonstrating robustness across modalities and noise conditions.

In summary, the proposed referenceless evaluation offers a *practical, efficient, and generalizable* alternative for assessing MT quality in realistic contexts. Future work will extend the framework to more language pairs and tasks, add semantic-level evaluation, and integrate end-to-end speech translation. By aligning evaluation with real downstream performance, this work advances a more holistic and reproducible paradigm for MT assessment.

Acknowledgment. This research is supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

10. Bibliographical References

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexandra Diaconu, Mădălina Vînaga, and Bogdan Alexe. 2026. Ro-n3ws: Enhancing generalization in low-resource asr with diverse romanian speech benchmarks.

Stefan Daniel Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of romanian bert. *arXiv preprint arXiv:2009.08712*.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 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.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. *arXiv preprint arXiv:2004.06063*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Iman Jundi and Gabriella Lapesa. 2022. How to translate your samples and choose your shots? analyzing translate-train few-shot cross-lingual transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *arXiv preprint arXiv:2107.10821*.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4):1006.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chi-kiu Lo and Samuel Larkin. 2020. Machine translation reference-less evaluation using YiSi-2 with bilingual mappings of massive multilingual language model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 903–910, Online. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*.

Mihai Alexandru Niculescu, Stefan Ruseti, and Mihai Dascalu. 2022. Rosummary: Control tokens for romanian news summarization. *Algorithms*, 15(12):472.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever.

2023. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W Oard, and Scott Miller. 2024. Translate-distill: learning cross-language dense retrieval by translation and distillation. In *European Conference on Information Retrieval*, pages 50–65. Springer.

Min Zhang, Xiaosong Qiao, Hao Yang, Shimin Tao, Yanqing Zhao, Yinlu Li, Chang Su, Minghan Wang, Jiaxin Guo, Yilun Liu, and Ying Qin. 2022. Target-side language model for reference-free machine translation evaluation. In *Machine Translation*, pages 45–53, Singapore. Springer Nature Singapore.

Min Zhang, Hao Yang, Yanqing Zhao, Xiaosong Qiao, Shimin Tao, Song Peng, Ying Qin, and Yanfei Jiang. 2023. Implicit cross-lingual word embedding alignment for reference-free machine translation evaluation. *IEEE Access*, 11:32241–32251.

Vilém Zouhar and Ondřej Bojar. 2024. Quality and quantity of machine translation references for automatic metrics. *arXiv preprint arXiv:2401.01283*.