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

Anonymous ACL submission

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. We compare four prominent MT tools by using them to translate tasks from Romanian into English and investigate their impact on text summarization, sentiment analysis, and authorship identification. Our findings demonstrate that while translation significantly boosts performance in summarization and sentiment analysis, it adversely affects the identification of authorship in poetry. In response to the observed performance disparities among MT models, we have developed a ranking system that aligns closely with human preferences. This system avoids reliance on professional ground-truth translations, which are typically required by traditional MT evaluation metrics like BLEU but can be biased by the quality of the reference and the translator's proficiency. Our approach provides a more authentic measure of MT quality, reflecting more accurately how these models perform in practical applications.

1 Introduction

Reliable evaluation of machine translation (MT) methods is a well researched topic of the past years, but challenges such as noise introduced by source and reference text quality or by human translators preferences, expensive acquisition of human translation references, and scarcity of varied enough evaluation datasets are yet to be overcomed.

The overwhelming majority of evaluation algorithms used for MT are based on ground-truth references, as noted by Lee et al. (2023), to either compute a similarity metric between the MT output and the reference or some kind of quality or ranking measure which after aggregated at dataset level can be used to compare the performance of different models.

As the main use case of evaluation metrics is to provide an insight for researchers about which MT model can be considered *better* at the system level rather than on individual translation instances, this creates the incentive for evaluation methods which do not require human-made translation references.

This paper explores the idea of ranking machine translation models from a source to a target language using generic mono-lingual NLP datasets by quantifying the performance impact on solving certain NLP tasks after translating the datasets into the target language. We apply our setting on the Romanian (source) to English (target) language pair as it was not studied comprehensively in translate-train settings and because Romanian is a relatively low-resource language.

The main contributions of the paper are:

- We study the translate-train technique on the language pair of Romanian to English, as it was never studied before, with three NLP tasks: sentiment analysis, text summarization, and authorship detection in poetry and discuss which use cases could benefit from a performance improvement and which could not;
- We propose an evaluation method for machine translation models which does not rely on ground-truth references and only needs generic NLP datasets in the source language;
- We report the results and their correlation with human judgement of our evaluation method on four popular translators: ChatGPT 3.5 Turbo, DeepL, Google Translate, and Mistralx7B Instruct v0.2.

Structure of the paper. The paper is organized as follows: in Section 2 we describe the required theoretical elements to properly understand and contextualize the paper, in Section 3 we present other referenceless evaluation methods, in Section

4 we propose our novel method for MT reference evaluation evaluation, in Section 5 we describe the experimental setup and results of our method on multiple datasets, in Section 6 we discuss the obtained results and their relation with human judgement, and finally Sections 7 and 8 present the limitations and main takeaways of the paper.

2 Background

In this section we describe a few theoretical notions with which the reader should be familiar with for an easier understanding of the work.

Human evaluation of machine translation is a wide topic consisting of a multitude of methodologies for assessing the quality of the output of MT systems in a measurable and systematic way. The study of Freitag et al. (2021) highlights the most common approaches found among methodologies: annotators providing scores on a discrete or continuous scale at segment or document level for various qualities of the test, identifying or rating mistakes and errors of multiple kinds from syntax, punctuation or wording, or less popular ones that use gap-filling or reading comprehension to evaluate the quality. The authors also note that many scale-based methodologies suffer from high variability induced by annotator's subjectivity.

Reference-based MT evaluation refers to the requirement of a ground-truth translation, usually crafted by a professional human translator, in order to provide an evaluation metric for a given MT system. We observe that the large majority of the metrics mentioned by Lee et al. (2023) are referenced-based.

Quality Estimation (QE) as a metric is a concept introduced at WMT19, as described by Fonseca et al. (2019), which puts forward the idea of using referenceless evaluation techniques for MT systems as inspired by QE approaches which historically revolved around estimating how *good* a given text is according to linguistic criteria.

Translate-train is a popular technique used to boost the performance of machine learning models on NLP tasks where models trained on the language at hand suffer from data scarcity. Works of Jundi and Lapesa (2022), Artetxe et al. (2023) or Jundi and Lapesa (2022) explore the advantages and scalability of this technique on various language pairs. When using translate-train, we usually have a source language which is low-resource and a target language which is high-resource and we translate the dataset at hand from the source to target language. Afterwards, we find-tune a mono-lingual or multi-lingual model to solve our desired NLP task. Other works such as the one presented by Yang et al. (2024) use translate-train in a knowledge distillation setting to train dual language encoders from mono-lingual language encoders.

3 Related work

Shortcomings of automated metrics are well-researched in the literature and serve as a central argument for metrics which are not solely based on similarity between outputs and references. WMT22 results published by Freitag et al. (2020) observe a low correlation between automatic metrics and human evaluation on three language pairs with varied structure. They highlight that neural-based evaluations such as COMET introduced by Rei et al. (2020) are superior to classical match-based evaluation. Both Mathur et al. (2020) and Reiter (2018) find an unstable behaviour of match-based metrics similar to BLEU in evaluating high-quality MT models and they are unreliable for comparing performance in pairwise settings. As noted by Kocmi et al. (2021), the use of inappropriate metrics held back the development of better translation systems for the past years.

Impact of reference quality on the correlation between automatic metrics and human judgement is studied by Zouhar and Bojar (2024) by acquiring four groups of translators with varying language expertise (identified by R1 to R4 with increasing expertise) and measuring how the scores of various metrics change across references created by groups. The surprising finding is that the best correlation with human judgement is found at the group R3 which is not the highest expertise group. This is against the common understanding that higher quality reference serve a better correlation with evaluation metrics. They also outline that a larger number of references outweighs the advantage of having few but very high quality references.

YiSi-2 with Bilingual Mappings, presented by Lo and Larkin (2020), is a referenceless evaluation metric that leverages bilingual mappings of massive multilingual language models. YiSi-2 evaluates MT quality by computing cross-lingual semantic similarity using pretrained multilingual models like BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). They found that projecting source embeddings into the target embedding

space using cross-lingual linear projection significantly improved correlation with human assessments. This approach addresses the language clustering effect observed in multilingual embeddings, thereby enhancing the metric's accuracy in evaluating translation quality across different languages.

Target-Side Language Model, proposed by Zhang et al. (2022), is a metric based on a target-side language model for reference-free MT evaluation. It evaluates translations by calculating sentence perplexity using a multilingual model like XLM-R. Their experiments on WMT19 datasets demonstrated that this approach is highly competitive, achieving strong correlations with human judgments at both segment and system levels. By focusing solely on target language fluency, this method simplifies the evaluation process and reduces dependency on source language complexities.

Implicit Cross-Lingual Word Embedding Alignment, introduced by Zhang et al. (2023), is a method that implicitly aligns cross-lingual word embeddings through multilingual knowledge distillation. This technique aligns sentence embeddings of parallel texts, resulting in better cross-lingual word embedding alignment. They incorporate this alignment into BERTScore and Word Mover's Distance metrics, achieving competitive results in reference-free MT evaluations. This approach highlights the effectiveness of using sophisticated embedding alignments to capture semantic equivalence between source and target texts without direct references.

COMET-QE developed by Rei et al. (2020), is a neural framework for reference-free MT evaluation. COMET-QE encodes segment-level representations of both source and translated texts and feeds them into a regressor to predict quality scores. This model benefits from fine-tuning on human-annotated quality estimation datasets, allowing it to learn nuanced quality signals that correlate well with human evaluations. This neural approach leverages advances in deep learning to provide robust and scalable quality estimation.

4 Proposed referenceless evaluation

We propose an evaluation method which is able to compare the performance of different MT models by measuring their impact on the performance of transformer-based language models on supervised NLP tasks after applying the translate-train approach.

The proposed method aims to remove the variability of results induced by the quality of references and reduce the cost of data acquisition necessary for comprehensive evaluation, while still maintaining a good correlation with human judgement.

Let's consider a fixed language pair (source $\rightarrow$ target) and a list of n machine translation models, each denoted with $T_i,\ i \in \overline{1,n}$, which should be compared against each other.

Instead of the usual translation pair dataset with texts in the source language and reference translations in the target language, we select m datasets in the source language, denoted with $\mathcal{D}_i^s,\ i \in \overline{1,m}$, which have supervised NLP tasks associated with them, such as any kind of text classification, summarization etc. For each NLP task, a bounded evaluation function such as ROUGE or F1-score should be available for measuring performance. For simplicity's sake, let's consider only one NLP task per dataset and denote the evaluation function associated with it $f_i(y, \hat{y})$, where y is a vector of ground-truth instances and $\hat{y}$ is a vector of predictions. For a simpler notation, we consider that the score associated with a given model $\mathcal{M}$ on a given dataset $\mathcal{D}_i^s$ to be $f_{\mathcal{D}_i^s}(\mathcal{M}) = f_i(\mathcal{M})$.

For each pair of datasets and MT models, we compute its translated version to the target language. Shortly, for the pair $(\mathcal{D}_i^s, T_j)$ we have the translated dataset $\mathcal{D}_{ij}^t = T_j(\mathcal{D}_i^s)$.

For each dataset, we select a pair of pre-trained transformed-based language models, one for the source language and one for the target language. It's preferred that the two models share the same architecture and number of parameters and differ only with respect to the weights. Thus, the performance variation induced by the architecture is minimized. We denote this pair of models with $(\mathcal{M}_i^s, \mathcal{M}_{ij}^t)$ for a given $\mathcal{D}_i^s$ dataset and corresponding translated version $\mathcal{D}_{ij}^t = T_j(\mathcal{D}_i^s)$.

Now, consider the proposed evaluation procedure for the MT models:

1. For each dataset $\mathcal{D}_i^s$, prepare a split suitable for training and evaluating the performance of a given model, such as a classic train/test split or a k-fold split;

2. Train and evaluate all source language transformers $\mathcal{M}_i^s$ on their associated datasets and tasks;

3. Train and evaluate all target language trans-

formers $\mathcal{M}_{ij}^t$ on all their translated datasets, namely the translate-train technique for each pair of datasets and MT models;

4. For each dataset $\mathcal{D}_i^s$ and available translator T_j, compute performance difference between the baseline transformer model in the source language and each model in the target language:

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

5. For each fixed MT model T_j, sum the differences between the baseline and translate-train performance across all datasets, and compute its final score:

$$S(T_j) = \Sigma_i \, \Delta(\mathcal{D}_i^s, T_j). \quad (2)$$

The performance differences described by Equation 1 represent the impact of the MT system output's quality on the NLP task and may be used to rank translators in a specific domain capability. Summing all the performance differences as described by Equation 2 should provide an accurate relative-ranking score for MT systems which should benefit from using more datasets.

5 Translate-train on Romanian to English

In this section we dive deep into the technical aspects of the translate-train technique as a part of the proposed referenceless evaluation method. We describe the three Romanian datasets used and their associated NLP tasks and evaluation metrics. We also discuss the usage of the four MT models used: ChatGPT3.5 Turbo [1], DeepL [2], Google Translate [3] and Mistralx7B Instruct v0.2 (Jiang et al., 2023) and, the training and evaluation setup of each experiment, and finally the results and their implications. The entire code base for the training setup will be made available as a supplement to the paper.

5.1 Datasets

Selection criteria. We reviewed available Romanian datasets associated with a plethora of different NLP tasks such as sentiment analysis, text summarization, fake news detection, dialect identification, named entity recognition, and others. The selection was based on the two main criteria: the dataset should have an associated NLP task solvable after translating to English, and the dataset should have more than ten thousands samples. Due to the first criterion, datasets with tasks such as named entity recognition were not selected, because mapping the labels from Romanian to English is non-trivial. We ended up by selecting three datasets described in the following paragraphs.

RoSent (Dumitrescu et al., 2020) is a movie and product reviews dataset with 28,000 samples, collected from unspecified web sources, which was manually annotated with positive or negative labels regarding the sentiment it communicates. Unfortunately, there is no information available about the data acquisition and labeling procedure provided by the authors. A stratified by label random subsample of 4,000 instances was selected for experiments.

RoTextSummarization (Niculescu et al., 2022) is a dataset consisting of new articles scraped from the Romanian news websites in the period of 2020 to 2022. The dataset contains around 72,000 articles alongside their summaries. A subsample stratified by the genre of each article was selected consisting of 8,000 instances.

Rupert [4] is Romanian poetry dataset with literary works of classic to contemporary authors with over 17,000 samples. The corpus contains over 500 authors, some of which have only one poem in the dataset, thus we decided to keep only the first 25 authors with respect to the number of poems they written. As the dataset is small and texts usually short, we selected a larger subsample percentage-wise of 5,000 instances stratified by each text's author.

Subsampling. To accommodate for our limited computational budget, we decided that for each dataset a random subsample of roughly 10 to 30% of the data, stratified where classes are present. Another cost taken into account was the cost of translation when using private translation models as they are quite expensive.

5.2 Tasks and metrics

Sentiment analysis was the associated task of RoSent. For each instance, we should predict the perceived positive or negative sentiment by a supposed reader. There was no class imbalance and we decided to choose F1-Macro score as our evaluation metric.

[1] https://openai.com/index/chatgpt/
[2] https://www.deepl.com/translator
[3] https://cloud.google.com/translate

[4] The Rupert dataset is available at https://huggingface.co/datasets/littlewho/Rupert.

Text summarization was the associated task of RoTextSummarization dataset and requires generating a text sample as close as possible to the reference summary of an article. The evaluation metric was chosen to be the popular ROUGE-L metric as it is widely used by summarization studies.

Authorship identification was selected as the associated NLP task for Rupert. This can be considered a classic text classification problem, but where stylistic features of the text at hand matter more than in usual scenarios. We also chose the F1-Macro score as the evaluation metric of this task.

5.3 Machine translation models

DeepL and **Google Translate** are one of the most used MT models on the market and they are specialized to provide high quality translations on a large number of language pairs. Both models were used via their API to translate the selected dataset subsamples at a cost of 20 euros, for each 500,000 characters. Each text was translated individually, not concatenated into a batch.

ChatGPT3.5 Turbo introduced by OpenAI and **Mistralx7B Instruct v0.2** introduced by Jiang et al. (2023) are both LLMs which can be successfully used in translation with decent results as noted by Kocmi et al. (2023). ChatGPT3.5 Turbo was instructed via a simple zero-shot prompt: *Translate from Romanian to English: <source text>* and it was entirely compliant and shown no significant hallucinations. On the other hand, due to computational constraints, Mistralx7B Instruct v0.2 was ran in a 16-bit quantized mode and required a few sentences of pre-programming to reduce hallucinations and non-conforming output formats: *You 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>*. Using the engineered prompt we obtained only 0.5% non-conforming outputs with hallucinations. We did not quantify textual hallucinations that did conform with the output format.

5.4 Transformer models

BERT was used for the two classification tasks of sentiment analysis and authorship identification in two variants, the Base Cased BERT introduced by Devlin et al. (2019), with English pre-training on a corpus of around 60 GiB , and the Romanian Cased BERT, which is a fine-tuned version of the base with additional pre-training on a Romanian language corpus of around 15 GiB by (Dumitrescu et al., 2020). The selected model has 110 million trainable parameters.

BART was used for the text summarization task, because it is a sequence-to-sequence model out-of-the-box. We used the base variant introduced by Lewis et al. (2020) which was pre-trained ona 160 GiB corpus, and a public Romanian variant [5] which was pre-trained from scratch on a 50 GiB Romanian corpus. Both models used have 140 million trainable parameters and a maximum output size of 1024 tokens.

The multilingual variant of BERT was also used to provide an additional interesting baseline for our experiments. Unfortunately, we did not find a multi-lingual variant of BART, so we proceeded with the Flan-T5-Small model which was developed by Chung et al. (2022).

5.5 Training setup

Training data for each NLP task was split into 5 folds and each training experiment was performed in a leave-one-out manner. Take note that the same folds were used in both the Romanian-only training and in the translate-train setting with the texts translated to English, so the results could be compared.

The texts in the training data were tokenized and trimmed to a number of tokens which represents the 95th length percentile across the entire dataset to reduce the computational resources needed to perform the experiments. In general, the 95th percentile trimming guaranteed a reduction of dataset size of almost 50%.

We used the suitable transformer model for each task as described above and fine-tuned it for 10 epochs with all layers not frozen. We made sure that the number of epochs is enough for the reported loss to converge for all experiments. We used the Huggingface framework to perform training with its default parameters, which in our case came in the form of AdamW optimizer with a learning rate of $5 \cdot 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and no weight decay. The batch sizes were 8 for the RoTextSummarization dataset, 16 for the Rupert dataset, and 32 for the RoSent dataset. The batch

[5] The Romanian BART model can be found at the following address https://huggingface.co/Iulian277/ro-bart-1024.

	RoSent		RoTextSummarization		Rupert	
	f1-macro	baseline diff	rouge-l	baseline diff	f1-macro	baseline diff
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**
GTranslate	91.82±00.49	+00.57	30.32±00.27	+06.06	64.88±01.98	-06.18
DeepL	**92.35±00.98**	**+01.11**	29.23±00.51	+04.97	63.66±03.00	-07.40
Mistralx7B	90.34±00.49	-00.91	28.08±00.64	+03.82	58.17±01.28	-12.89
Romanian results						
Romanian	91.24±00.30	-	24.26±00.43	-	71.06±01.19	-
Multilingual	87.61±00.81	-03.63	24.33±00.26	+00.07	62.32±01.67	-08.74

Table 1: 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.

sizes were selected such that the available VRAM is entirely used.

All training and evaluation was performed in single-GPU setups on NVIDIA GeForce RTX 4090 GPUs consisting of about 150 hours of total compute time for all the experiments.

5.6 Results

In Table 1 we report the results of the experiments on all datasets with Romanian, multilingual, and English transformers paired with each translator. The *baseline diff* column contains the results as described by Equation 1 in our proposed evaluation procedure for the MT models. In the following paragraphs, we discuss the main takeaways.

Performance boost on basic tasks. We observe that the translate-train approach improved the results in 3 out of 4 cases for the sentiment analysis task on RoSent, and in all cases for the text summarization task on RoTextSummarization. We acknowledge that results on the summarization task may be overly optimistic because we translated also the ground-truth summaries, and due to the behaviour of ROUGE-L metric on the Engish language.

The findings are consistent with investigations on other languages such as those presented by Jundi and Lapesa (2022), who performed experiments on 14 languages: Arabic, Bulgarian, German, Greek, Spanish, French, Hindi, Russian, Swahili, Thai, Turkish, Urdu, Vietnamese, and Chinese. We conclude that Romanian NLP tasks could benefit of performance improvements on tasks which do not require complex textual competency, a finding which was confider on other languages by (Artetxe et al., 2023).

Performance degradation on poetry. No translator was able to improve the performance of authorship identification in the translate-train settings. We hypothesize that the stylistic features specific to poems are of great importance for solving this NLP task, and those features are most likely lost in translations. Similar findings are described again by Artetxe et al. (2023), the authors explaining how tasks which require a *deeper* textual understanding have a hard time benefiting out of the translate-train method.

The MT system which performs the best varies across the NLP tasks. We interpret this as different MT systems having strong and weak points in separate capabilities.

6 Results of MT models evaluation

In Table 2, we report the scores assigned to each MT system according to Equation 2 as described earlier in our evaluation procedure. Their scores are the sum of the NLP task performance difference between each MT system in the translate-train setting and the baseline result provided by the Romanian transformer. Detailed results of the translate-train experiments were presented in the earlier Table 1.

In the following paragraphs, we present how we collected human judgements for a subsample of translations taken from the three NLP datasets we used. We used those judgements to analyse the correlation between our referenceless evaluation methodology and human assessment.

6.1 Human-judgement data acquisition

To validate that our evaluation method is in line with human judgement we decided to collect human feedback in the form of translation preferences between alternative translations generated by different MT systems. Each user was presented with a series of questions, each question presenting a Ro-

Translator	Our score
ChatGPT3.5	+01.37
GTranslate	+00.45
DeepL	-01.32
Mistral7xB	-09.98

Table 2: Cumulative scores for all MT systems by considering their performance difference between the translate-train result and the baseline result in Romanian language.

manian text taken from one of our NLP datasets and two alternative translation to English of the given text. The translations were generated with one of the four evaluated MT systems and the system used was not disclosed to the user. Text samples and the order in which translations were shown to the user were randomized to reduce the risk of bias. Each user had to use a slider to choose which alternative translation prefers, choices ranging from left/right being slightly or much better, or choosing a tie between the two.

Each user was presented a questionnaire of 60 texts randomly sampled of a pool of 240 texts selected for evaluation. Each user received its own random version of the questionnaire to reduce the possible bias by a certain selection of texts.

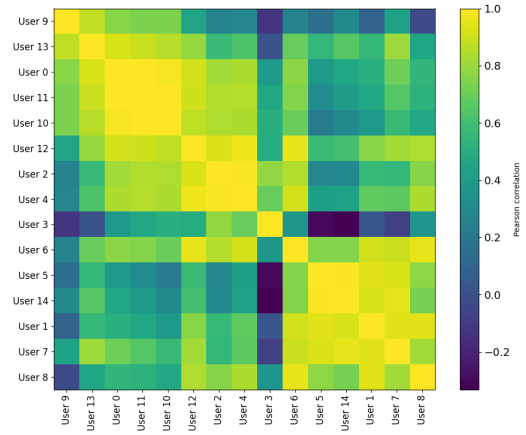

Figure 1: Matrix of pairwise Pearson correlation coefficient between the votes the questioned users. The columns were re-ordered to highlight correlation clusters.

We successfully collected about 900 judgements from 15 volunteering students and professors specialized in the fields of computer science, law, and foreign languages. All respondents were native Romanian speakers with an estimated language competency of at least B2 for English.

To rank the evaluated MT systems based on the votes, we used a simple scoring scheme: both systems received 0.5 points if the user voted *tie* or 1 full point to the system which received a favourable vote. We also tried having 0.5/1.0/1.5 scoring scheme as our votes had two possible values for favouring one system over the other, but we did not see any significance in the results we present below, so we stuck with the simpler scoring scheme. The results after aggregating the user votes can be observed in Table 3 along our automatic evaluation to highlight the agreement between the two.

6.2 Correlation between our evaluation and human judgement

To assure the reader of the consistency of human judgements across different persons, in Figure 1 we present a heatmap of pairwise Pearson correlation coefficients between the votes of the users. We observe the tendency of cluster formation and the existence of at least one outlier, namely *User 3*, but overall the results present a positive correlation.

In terms of descriptive statistics for the correlations above, we report a mean of 0.6159, a median of 0.6926, and a standard deviation of 0.3001. The interquartile range (IQR) contains values between 0.4391 and 0.8455, which further emphasizes a moderate to strong correlation between human judgements. The standard deviation is expected to be high for human answers.

In Table 4, we present the Pearson correlation between the results of our proposed evaluation metric and the results of human judgements for all datasets and at the system level aggregating all results. We observe a strong correlation at the system level of 0.88, which should improve by selecting more and higher quality datasets.

The lowest correlation observed is for the RoSent dataset, which may suggest that datasets of lower complexity are worse for evaluating high-quality MT systems. This is consistent with the fact that the dataset with the highest textual complexity, Rupert, which contains poetry, it's best correlated with human judgement. We may conclude that *harder* datasets better estimate how well a MT system behaves.

7 Limitations

In spite of the high correlation with human judgement of our proposed evaluation method, we must acknowledge a series of limitations of our approach which we hope to overcome in future works.

Translator	RoSent	RoTextSummarization	Rupert	Total	Our evaluation
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 3: Scores allotted to each MT system by using the 0.5p/1p scoring scheme for user votes. The last column contains the scores proposed by our evaluation strategy to highlight the agreement with human judgement.

Dataset	Pearson Correlation
RoSent	0.4756
RoTextSummarization	0.6423
Rupert	0.9769
All	0.8741

Table 4: Pearson correlation between the human judged scores and our proposed referenceless evaluation method. Results are presented per-dataset and for all datasets combined.

Only one language pair. In all the experiments we used only one language pair in only one direction, Romanian to English, which may pose a threat to the generality of the proposed methodology. The results may not reproduce in scenarios of high-resource to high-resources or high-resource to low-resource languages.

Small pool of human evaluators. The number of volunteers we used to judge translations is small and may not be representative for larger populations. We also acknowledge as highlighted by other works such as (Freitag et al., 2021), that crowd-sourced volunteers may not output judgements in line with the judgements of professional translators.

Translate-train is limited. Trainslate-train is limited to a certain subset of NLP tasks, some of them being harder or almost impossible to solve in such a settings, such as named-entity recognition (NER) or question answering. This reduces the available datasets that can be used and also the capabilities of the MT systems that can be evaluated with our approach.

Relative ranking is not entirely accurate. Our approach tends to rank MT systems much closer to each other than the human judgements, which may suggest that we cannot properly distinguish between the performance of high-quality MT systems. This may be due to dataset quality and complexity or the approach in itself.

8 Conclusions

We presented a novel referenceless evaluation method for assessing machine translation models by leveraging their performance impact in translate-train settings across various natural language processing tasks. By translating Romanian texts into English and subsequently evaluating the impact on text summarization, sentiment analysis, and authorship identification, we demonstrated significant improvements in the first two tasks while noting a performance decline in the latter. This highlights that while MT can enhance certain NLP applications, tasks requiring nuanced textual comprehension, such as poetry authorship identification, may suffer from translation-induced distortions.

The proposed evaluation approach circumvents the need for professional ground-truth translations, traditionally required by metrics like BLEU, thus reducing potential biases and costs. By relying on generic NLP datasets in the source language, this method provides a more authentic reflection of MT model performance in practical applications. The findings also revealed variability in MT model performance across different NLP tasks, suggesting that different MT systems possess distinct strengths and weaknesses depending on the domain of the text. The developed ranking system, which aligns closely with human preferences, further validates the robustness and reliability of the proposed referenceless evaluation method.

In conclusion, the study's innovative approach offers a cost-effective alternative for MT model evaluation, addressing some of the longstanding challenges in the field. While the methodology showed high correlation with human judgments, especially in more complex datasets, it also highlighted the need for further exploration across diverse language pairs and additional NLP tasks to fully generalize its applicability. This work paves the way for more accurate and scalable MT evaluations, fostering advancements in machine translation technologies.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

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.

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*.

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. *arXiv preprint*.

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, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. 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*.

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*.

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

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. *arXiv preprint*.

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*.