Trained MT Metrics Learn to Cope with Machine-translated References

Jannis Vamvas[2*] **Tobias Domhan**[1] **Sony Trenous**[1]
Rico Sennrich[2] **Eva Hasler**[1]
[1]Amazon AI Translate, Berlin
[2]University of Zurich
{vamvas,sennrich}@cl.uzh.ch, {domhant,trenous,ehasler}@amazon.com

Abstract

Neural metrics trained on human evaluations of MT tend to correlate well with human judgments, but their behavior is not fully understood. In this paper, we perform a controlled experiment and compare a baseline metric that has not been trained on human evaluations (*Prism*) to a trained version of the same metric (*Prism+FT*). Surprisingly, we find that Prism+FT becomes more robust to machine-translated references, which are a notorious problem in MT evaluation. This suggests that the effects of metric training go beyond the intended effect of improving overall correlation with human judgments.

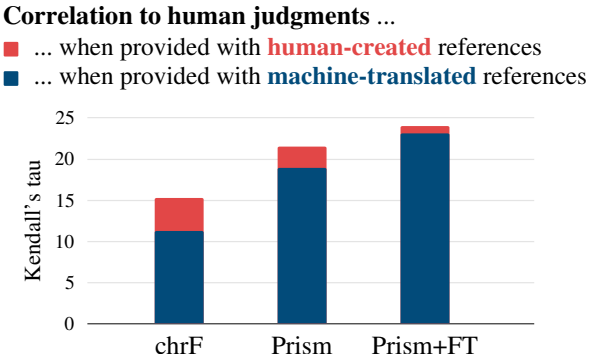

Figure 1: Metrics for MT quality have a lower segment-level correlation with human judgments when provided with machine-translated references. However, trained metrics, such as our Prism+FT, become more robust to the use of machine translations as references.

1 Introduction

While trained evaluation metrics for machine translation (MT) tend to have a high correlation with human judgments (Freitag et al., 2022b), they remain black boxes, sometimes behaving in unexpected ways (Amrhein and Sennrich, 2022; Rei et al., 2023). This calls into question whether a metric's utility can be measured solely by its correlation with human judgments.

In this paper, we intentionally provide MT metrics with *machine-translated reference translations*, as opposed to human-created references, and investigate how this factor influences the behavior of a metric. In MT evaluation research, the human translators who create reference translations are usually asked to produce them from scratch, in order to avoid references that are machine-translated or post-edited (Kocmi et al., 2022). Nevertheless, traces of MT have been detected in some reference sets (Kloudová et al., 2021; Akhbardeh et al., 2021; Kocmi et al., 2022). It is therefore important to understand how metrics behave under such references.

In our experiments, we use a surrogate for real post-edited references in the form of error-free output by various systems from the WMT 2021 news translation task (Akhbardeh et al., 2021). Our results show that there is a stark difference between trained and non-trained metrics: While trained metrics maintain most of their accuracy when provided with such MT-derived references, non-trained metrics exhibit a substantial drop in accuracy.

To corroborate this observation, we perform a controlled experiment involving Prism (Thompson and Post, 2020), a metric that is based on a multilingual MT system. The original version of Prism can be considered non-trained, since it learns from parallel sentences without human judgments. We then fine-tune Prism on a dataset of human judgments, using a bidirectional pairwise ranking approach.

As expected, the segment-level correlation of Prism increases during fine-tuning, indicating that the metric learns to better predict human judgments (Figure 1). Moreover, we find that fine-tuning narrows the gap in performance between human-created and machine-translated references. Our experiment thus indicates that training a metric on human evaluation data can influence its behavior in a way that is not captured by global correlation with human judgments. Code to reproduce our

[*]Work done during an internship at Amazon.

findings will be made available.[1]

To summarize, the paper makes the following contributions:

- We propose a metric evaluation setup that intentionally uses machine-translated references, and demonstrate that non-trained metrics perform poorly in this setup.
- We present an approach for fine-tuning Prism on human judgments that significantly improves segment-level correlation on unseen test data.
- We show that fine-tuning Prism on human judgments makes it more robust to the use of machine-translated references.

2 Background

2.1 Reference-based Evaluation

Automatic evaluation of MT is often performed by comparing the system output with one or more reference translations, using an evaluation metric. Evaluation metrics can be roughly divided into *trained* and *non-trained* metrics. Trained metrics receive supervision from human judgments of past machine translations. For example, Sellam et al. (2020) and Rei et al. (2020; 'COMET') fine-tuned a pre-trained sentence encoder on such human judgments, using regression or ranking objectives.

Non-trained metrics, on the other hand, rely on a heuristic to make the comparison. Metrics such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015) are based on the overlap of words or characters between the system output and the reference. Thompson and Post (2020) use the perplexity of a neural sequence-to-sequence model, called Prism, that has been trained on multilingual MT. Systematic comparisons of evaluation metrics (Freitag et al., 2022b) have shown that trained metrics tend to correlate better with human judgments than non-trained metrics do, especially if the latter are based on overlap heuristics.

2.2 Quality of Reference Translations

The reliability of reference-based evaluation metrics also depends on the quality of the references they are provided with (Freitag et al., 2021b). A notorious source of noise in references is *translationese*, which is characterized by monotonicity with respect to the source sequence and a high n-gram overlap with system translations (Freitag et al., 2020). Freitag et al. (2020) have shown that translationese references cause BLEU scores to be higher, and the scores are dominated by matches of common, unspecific n-grams. They find that BLEU scores under non-translationese references tend to be lower, but more precise.

Agarwal et al. (2023) observed that post-edited references for spoken language translation seem to inflate BLEU scores, but not the scores of COMET. However, the relationship between metric training and the quality of reference translations has not been studied in detail. In this paper, we hypothesize that robustness to machine-translated references may partially explain why trained metrics are more accurate in practice.

3 Experimental Setup

3.1 Measuring Global Correlation

For measuring the overall correlation of a metric to human judgments, we follow the WMT 2021 metrics task (Freitag et al., 2021b) and use MQM annotations of submissions to the 2021 WMT news translation task (Akhbardeh et al., 2021). The evaluation data cover two domains, news and TED talks. Table A5 reports statistics for these data.

We closely replicate the methodology of the WMT 2021 metrics task. On the segment level, we report Kendall's tau coefficient across all segments and systems; on the system level, we report *pairwise accuracy* (Kocmi et al., 2021), i.e., the ratio of system pairs that a metric ranks in the same order as human annotators have. Following the shared task, we only consider system translations and exclude human translations from the evaluation. We then perform *perm-both* hypothesis tests (Deutsch et al., 2021) to validate metrics comparisons at $\alpha = 0.05$.

3.2 Measuring the Effect of Machine-translated References

In the context of our analysis, we use error-free system translations from the WMT 2021 news translation task as a surrogate for real post-edited references. Specifically, we randomly select system translations that have been annotated according to the MQM standard and in which no annotator has marked an error. This approach allows us to simulate a post-editing process without the cost and noise incurred by actual post-editing.

[1] https://github.com/amazon-science/prism-finetuned

Source sequence (English)
Face masks are mandatory across the state of California, even in fresh air.

Human-created reference (German)
Gesichtsmasken sind im ganzen Bundesstaat Kalifornien vorgeschrieben, auch im Freien.

Machine-translated reference (German)
Gesichtsmasken sind im gesamten Bundesstaat Kalifornien Pflicht, auch an der frischen Luft.

Figure 2: Example of a machine-translated reference compared to the standard reference created by a human translator. The machine-translated reference is more literal (*an der frischen Luft* 'in fresh air').

Figure 2 and Appendix F juxtapose some examples of error-free system translations and the standard, human-created reference translations. The former tend to be more literal and more aligned to the source, both in terms of syntax and content.

It should be noted that when we evaluate a metric in this analysis, we draw from the same set of systems and human annotations as we do for extracting the references. We take care to properly separate the system translations used as a reference from those that are evaluated based on that reference.

To calculate segment-level correlation, we sample a random error-free translation from an unrelated system, for each system output.[2] To calculate system-level pairwise accuracy, we use different sets of references depending on the pair of systems that is compared. Figure 3 shows that our approach is comparable to cross-validation. For every pair of systems that we consider when calculating the pairwise accuracy of a metric, we select one reference translation from an unrelated system, independently per segment. As a consequence, we use slightly different reference sets for ranking different pairs of systems.

We then compare the accuracy of a metric when provided with the machine-translated references to its accuracy when using the standard references. To ensure comparability, we skip all the segments where no machine-translated reference is available (which is either because the segment has not been part of the annotation study or because annotators have found an error in every system translation). The metric accuracies for both ref_{std} and ref_{mt} are

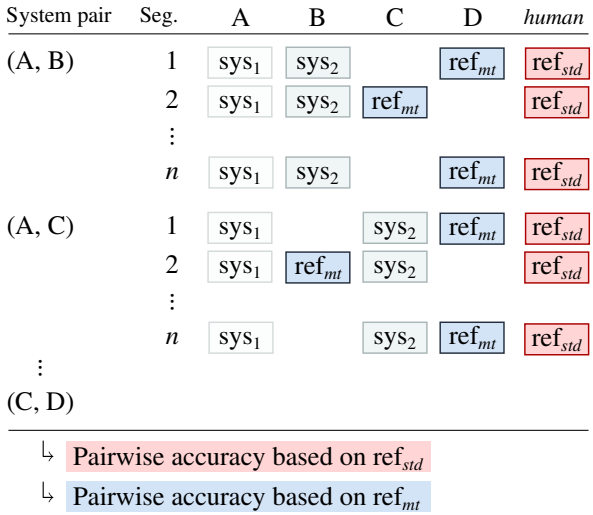

Figure 3: To measure the effect of machine-translated references, we use error-free output from other, unrelated MT systems as references. For example, when comparing system A to system B, we use a translation from either system C, D, etc. as a reference for each segment.

thus calculated based on a subset of the segments used to calculate global correlation. Table A5 shows that only for one language pair a substantial number of segments need to be skipped (Chinese–English news). For the other language pairs, between 0% and 4.5% of the segments are skipped.

4 Fine-tuning the Prism Metric

Prism (Thompson and Post, 2020) is a reference-based evaluation metric that relies on the paraphrasing probability between a system translation and a reference. The probability is estimated by a multilingual NMT model as a zero-shot translation direction. The model is expected to prefer mere copies of the source sequence to more creative paraphrases, which is especially useful for reference-based evaluation.

The NMT model uses the reference as a source

[2] Segment-level correlation is calculated jointly across all segments and systems, and as a consequence, using different references to evaluate the translations of different systems adds some noise to the correlation. However, we expect that the correlation is dominated by the segment axis and not by the system axis. Our findings on the segment level are consistent with our findings on the system level.

sequence x and the system translation as a hypothesis y, or vice versa. The segment-level score S is then calculated from token-level log-probabilities:[3]

$$S(y|x) = \frac{1}{|y|}\sum_{t=1}^{|y|}\log p(y_t|y_{i<t}, x).$$

By default, Prism uses the average of both paraphrasing directions:

$$\text{Prism}(\text{sys}, \text{ref}) = \frac{1}{2}S(\text{sys}|\text{ref}) + \frac{1}{2}S(\text{ref}|\text{sys}).$$

An overall score for a system can then be calculated as an average over a collection of segments.

4.1 Training Objective

In order to fine-tune Prism, we combine a standard cross-entropy objective and a bidirectional pairwise ranking objective.

For the *cross-entropy objective*, we use the source sequence (src) and the reference translation (ref) of the training examples to continue the cross-entropy training:

$$L_{\text{src}\to\text{ref}} = -S(\text{ref}|\text{src}).$$

Our goal in using this objective is to familiarize Prism with the segments to which the human judgments refer, and to prevent catastrophic forgetting during the fine-tuning stage.

In addition, we propose a *bidirectional pairwise ranking objective*. In the forward direction, we train Prism to correctly rank two system translations (sys^+ and sys^-), conditioned on the reference (*forward ranking*):

$$L_{\text{ref}\to\text{sys}} = \max\{0, \epsilon - S(\text{sys}^+|\text{ref}) + S(\text{sys}^-|\text{ref})\},$$

where ϵ is a margin value. We add a second ranking loss for the reverse paraphrasing direction, i.e., for reconstructing the reference from either of the system translations (*backward ranking*):

$$L_{\text{sys}\to\text{ref}} = \max\{0, \epsilon - S(\text{ref}|\text{sys}^+) + S(\text{ref}|\text{sys}^-)\}.$$

The complete fine-tuning objective is:

$$L = \alpha L_{\text{src}\to\text{ref}} + (\frac{1}{2}L_{\text{ref}\to\text{sys}} + \frac{1}{2}L_{\text{sys}\to\text{ref}}),$$

where α is a scalar to balance the two terms.

Figure 4 is a schematic illustration of the objectives for pre-training, fine-tuning, and inference.

[3]This score is called H in the original definition. We use S instead, to avoid confusion with cross-entropy (which is $-S$).

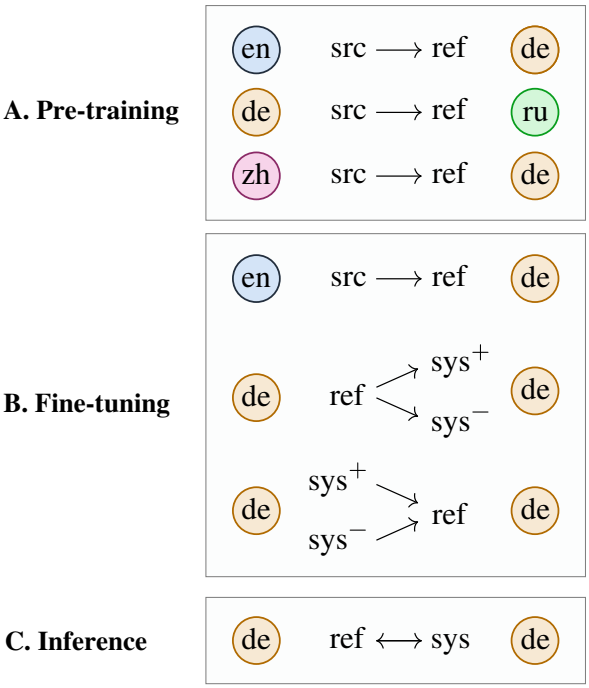

Figure 4: Schematic illustration of the sequences used for pre-training, fine-tuning, and applying the Prism model to MT evaluation. Prism has been **(A)** pre-trained on multilingual translation to and from 39 languages as described by Thompson and Post (2020); inference **(C)** makes use of the zero-shot paraphrasing capability acquired by the model during pre-training. We add a fine-tuning stage **(B)** with data derived from human evaluations of MT. In this illustration, Prism is fine-tuned on English–German examples.

4.2 Training Data

For fine-tuning Prism, we use human judgments of submissions to the 2020 WMT news translation tasks (Barrault et al., 2020), collected by Freitag et al. (2021a).[4] These annotations are based on the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014) and have been shown to correlate better with automatic metrics than previous direct assessments, especially when the evaluation concerns high-quality translations (Freitag et al., 2021a,b). Specifically, we train Prism on human judgments for English–German and Chinese–English translations of news. We train a single model jointly on both language pairs.

To use the human judgments for training on pairwise ranking, the direct MQM assessments need to be converted into relative rankings of translation pairs. In previous work, direct (non-MQM) assess-

[4]Submission data are available at https://github.com/google-research/mt-metrics-eval and the MQM annotations are available at https://github.com/google/wmt-mqm-human-evaluation

ments have been normalized and aggregated across annotators before being compared (Ma et al., 2019). Since MQM ratings are known to have low inter-annotator agreement on the segment level (Freitag et al., 2021b), we opt for intra-annotator pairing instead. Specifically, we only pair translations that have been rated by the same annotator, and we do not compare MQM scores across annotators. Relative rankings are created independently for each annotator and then concatenated. Furthermore, we only pair translations that have a score difference greater than 0.1, which would correspond to a minor fluency or punctuation error. Taken together, these criteria should ensure there is a noticeable difference between the quality of two system translations sys^+ and sys^- in the eyes of at least one annotator. We hold out 5000 relative rankings from the resulting training data as a validation set and use it to select hyperparameters. Detailed statistics for the training data are provided in Table A4.

4.3 Implementation Details

The fine-tuning was implemented in Fairseq (Ott et al., 2019). We start with the original `Prism39` model released by Thompson and Post (2020).[5] We then fine-tune the model for a single pass over the training data, using Adam. The initial learning rate is set to 1e-4 without any warm-up steps. We use half-precision training and an effective batch size of 360k tokens. Other settings match the pre-training setup of Prism.

We set the margin hyperparameter ϵ to 0.1, and the cross-entropy weight α to 0.1 as well. The hyperparameters have been selected based on segment-level correlation on the validation set. Since we jointly train on two language pairs, we iterate over batches for each language pair in a round-robin fashion, upsampling the smaller language pair. Fine-tuning takes about one hour on a `p3.8xlarge` AWS instance, which has 4 Tesla V100 GPUs with 16 GB of memory.

5 Results

Effect of fine-tuning Prism Table 1 shows that fine-tuning Prism has the intended effect: *Fine-tuning Prism on human judgments of machine translations significantly improves correlation with human judgments on an unseen test set.* The effect of fine-tuning is especially pronounced for the English–German and Chinese–English language pairs, since the metric was fine-tuned on those pairs. Interestingly, we also observe positive cross-lingual transfer to the English–Russian language pair, which was not seen during fine-tuning. Table 2 shows that the positive effect of fine-tuning extends to the TED Talks domain, even though the metric was not fine-tuned on this domain.

	EN–DE	EN–RU	ZH–EN
Prism	19.3	22.4	28.8
Prism+FT	**25.3**	**23.7**	**31.5**

Table 1: In-domain accuracy of Prism on WMT 2021 news translation submissions. We report segment-level Kendall's tau correlation to human judgments. Bold font denotes that the improvement achieved through fine-tuning is significant with $\alpha = 0.05$. Note that Prism+FT has not been fine-tuned on the EN–RU language pair.

	EN–DE	EN–RU	ZH–EN
Prism	24.2	21.9	19.6
Prism+FT	**26.9**	22.3	**21.9**

Table 2: Out-of-domain accuracy of Prism on WMT 2021 system translations of TED talks in terms of segment-level Kendall's tau. Bold indicates that the improvement is significant with $\alpha = 0.05$.

Effect of using machine-translated references Table 3 reports the segment-level correlation of different metrics when using either standard references or machine-translated references. Note that the values for Prism slightly differ from Tables 1 and 2 because this analysis is based on a subset of the segments. *We find that the correlation of metrics to human judgments tends to decrease under machine-translated references.* For the Chinese–English dataset the relative decline is smaller than average, but is still noticeable for most metrics.

In Table 3, when comparing the non-trained metrics (above the horizontal line) to the trained metrics (below the line), we observe that the decline in correlation is smaller for the trained metrics. An especially interesting comparison is between Prism and Prism+FT, given that the two metrics differ only in the training data. *Prism+FT is consistently more robust to machine-translated references than Prism, indicating that the metric learns to cope with such references during the fine-tuning stage.*

With respect to system-level pairwise accuracy (Table 4), we observe a similar trend.

[5] https://data.statmt.org/prism/

	EN−DE		EN−RU		ZH−EN		Average	
	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}
BLEU	8.4	7.0 **(-16.7%)**	12.1	11.8 **(-2.5%)**	15.2	14.8 **(-2.6%)**	11.9	11.2 **(-5.9%)**
chrF	11.1	8.3 **(-25.2%)**	19.3	13.8 **(-28.5%)**	16.7	15.7 **(-6.0%)**	15.7	12.6 **(-19.7%)**
Prism	18.9	18.2 **(-3.7%)**	22.4	20.6 **(-8.0%)**	24.2	23.5 **(-2.9%)**	21.8	20.8 **(-4.9%)**
Prism+FT	24.9	24.4 **(-2.0%)**	23.7	22.3 **(-5.9%)**	26.6	26.8 **(0.8%)**	25.1	24.5 **(-2.3%)**
COMET	25.1	24.6 **(-2.0%)**	27.6	25.4 **(-8.0%)**	32.1	32.1 **(0.0%)**	28.3	27.4 **(-3.2%)**

Table 3: Segment-level correlation of MT metrics when provided with the standard references (ref_{std}) of the WMT21 metrics news subtask (Freitag et al., 2021b), and with machine-translated references (ref_{mt}). The percentages denote the relative change in correlation when falling back to machine-translated references. The trained metrics, Prism+FT and COMET (`wmt21-comet-mqm`), have a more favorable relative change than the non-trained metrics, which indicates higher robustness to machine-translated references.

	EN−DE		EN−RU		ZH−EN		Average	
	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}
BLEU	89.7	74.4 **(-17.1%)**	70.3	58.2 **(-17.2%)**	61.5	61.5 **(0.0%)**	73.8	64.7 **(-12.4%)**
chrF	87.2	71.8 **(-17.7%)**	74.7	56.0 **(-25.0%)**	60.3	56.4 **(-6.5%)**	74.1	61.4 **(-17.1%)**
Prism	85.9	73.1 **(-14.9%)**	83.5	62.6 **(-25.0%)**	61.5	56.4 **(-8.3%)**	77.0	64.0 **(-16.8%)**
Prism+FT	89.7	80.8 **(-9.9%)**	80.2	61.5 **(-23.3%)**	61.5	61.5 **(0.0%)**	77.1	67.9 **(-11.9%)**
COMET	79.5	84.6 **(6.4%)**	68.1	65.9 **(-3.2%)**	60.3	55.1 **(-8.6%)**	69.3	68.5 **(-1.1%)**

Table 4: System-level pairwise accuracy of MT metrics when provided with the standard references of the WMT21 metrics news subtask (Freitag et al., 2021b), and with machine-translated references. Again, the trained metrics, Prism+FT and COMET (`wmt21-comet-mqm`), tend to be more robust to machine-translated references.

Prism+FT does not show significantly higher pairwise accuracy than Prism when using standard references, which is explained by the high statistical variance of the pairwise accuracy metric. But again, Prism+FT appears more robust to machine-translated references than Prism. Finally, Appendix B reports results for the TED talks domain, where the same patterns can be observed.

Ablation Study We perform an ablation study to measure the influence to the three terms in the Prism fine-tuning objective. Appendix A shows that removing either of the three terms decreases segment-level correlation. The ablation shows that the cross-entropy objective has the additional effect of stabilizing the model: Without cross-entropy, the average probability scores output by Prism shift from 0.47 to 0.35 after a single epoch of fine-tuning, and the BLEU achieved by the Prism translation model on an unseen test set clearly declines.

6 Related Work

Machine translations as references Popovic et al. (2016) first investigated the potential of using post-edited machine translations as references, finding that post-edited translations stemming from high-quality systems are better references than those from low-quality systems. Toral (2019) argued that post-edited machine translations can be seen as an exacerbated form of translationese (*post-editese*). Combined with the finding of Freitag et al. (2020) that translationese references are less favorable than intentionally paraphrased references, this suggests that machine translations, even if post-edited, are a challenge for MT evaluation.

Albrecht and Hwa (2007) propose to train an evaluation metric using non-annotated translations of other systems as *pseudo-references*. They hypothesize that a metric can learn to detect and to constructively utilize any errors in these references. Yoshimura et al. (2019) instead use a paraphrase identifier to filter pseudo-references based on their paraphrastic similarity to a human-created reference. Finally, minimum Bayes risk decoding (Kumar and Byrne, 2004) employs pseudo-references for generating translations, and has been shown to depend on robust metrics as well (Freitag et al., 2022a; Amrhein and Sennrich, 2022).

Training a sequence-to-sequence model on pairwise ranking Pairwise ranking has commonly been used to train SVM (Ye et al., 2007; Duh, 2008; Stanojević and Sima'an, 2014) and neural network encoders (Guzmán et al., 2015; Dušek et al., 2019). A more recent approach has been to fine-tune pre-trained sentence encoders so that the embedding similarities of two hypotheses and the reference and/or source are optimized for pairwise ranking (Rei et al., 2020; Zhang and van Genabith, 2020), in which case the max-margin loss reduces to a triplet margin loss (Schroff et al., 2015). In this paper, we do not rely on the similarity of sentence embeddings but use the perplexity of a sequence-to-sequence model as a metric.

Since we optimize perplexity given positive and negative examples, our fine-tuning approach becomes very similar to contrastive learning for NMT. Typical applications of contrastive learning try to eliminate specific translation error types by creating perturbed versions of the training references (Yang et al., 2019; Hwang et al., 2021). A similar objective has been used for discriminative re-ranking of translation candidates (Shen et al., 2004; Yu et al., 2020). In this paper, however, the goal is not to improve translation output but to train an evaluation metric on human judgments.

7 Conclusion

We have shown that metrics without supervision by human judgments, such as BLEU and chrF, tend to be inaccurate under machine-translated references, while trained metrics are more robust. In order to methodically examine this phenomenon, we have trained the Prism evaluation metric on a dataset of human judgments. Our experiments show that fine-tuning improves the segment-level accuracy of Prism on an unseen test set across multiple language pairs and domains, and clearly increases its robustness to machine-translated references.

One conclusion to draw from our findings is that post-edited references likely diminish the accuracy of reference-based metrics and should be avoided. A second conclusion is that if it cannot be ruled out that references originate from MT, as is often the case in practice, trained metrics are to be preferred. Fine-tuning a metric such as Prism on reference-based evaluation can thus be seen as a technique to let the metric make the best out of reference translations in the wild.

Limitations

Our study is mainly limited by the data we use for fine-tuning and evaluating Prism. The experiments are based on three language pairs only. Automatic MT evaluation is relevant for many more language pairs and language families, including and maybe especially so for low-resource settings.

Secondly, it should be mentioned that the machine translations we use in our analysis have been generated by systems based on a similar technology. Almost all of the systems seem to use the Transformer architecture, and they have all been trained on similar data (Akhbardeh et al., 2021). It is possible that our findings do not generalize to the evaluation of other varieties of MT, such as rule-based systems, or to reference-based evaluation metrics that use large language models (Kocmi and Federmann, 2023).

Acknowledgements

We thank Bill Byrne, Felix Hieber, Brian Thompson and Ke Tran for comments on an earlier stage of this project. JV and RS acknowledge funding by the Swiss National Science Foundation (project MUTAMUR; no. 176727).

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 evaluation campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa,

Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Joshua Albrecht and Rebecca Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 296–303, Prague, Czech Republic. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194, Columbus, Ohio. Association for Computational Linguistics.

Ondřej Dušek, Karin Sevegnani, Ioannis Konstas, and Verena Rieser. 2019. Automatic quality estimation for natural language generation: Ranting (jointly rating and ranking). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 369–376, Tokyo, Japan. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. 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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 805–814, Beijing, China. Association for Computational Linguistics.

Yongkeun Hwang, Hyeongu Yun, and Kyomin Jung. 2021. Contrastive learning for context-aware neural machine translation using coreference information. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1135–1144, Online. Association for Computational Linguistics.

Věra Kloudová, Ondřej Bojar, and Martin Popel. 2021. Detecting post-edited references and their effect on human evaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 114–119, Online. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp

Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

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. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popovic, Mihael Arčan, and Arle Lommel. 2016. Potential and limits of using post-edits as reference translations for MT evaluation. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 218–229.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada. Association for Computational Linguistics.

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)*, pages 2685–2702, Online. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha, Qatar. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Yang Ye, Ming Zhou, and Chin-Yew Lin. 2007. Sentence level machine translation evaluation as a ranking. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 240–247, Prague, Czech Republic. Association for Computational Linguistics.

Ryoma Yoshimura, Hiroki Shimanaka, Yukio Matsumura, Hayahide Yamagishi, and Mamoru Komachi. 2019. Filtering pseudo-references by paraphrasing for automatic evaluation of machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 521–525, Florence, Italy. Association for Computational Linguistics.

Lei Yu, Laurent Sartran, Po-Sen Huang, Wojciech Stokowiec, Domenic Donato, Srivatsan Srinivasan, Alek Andreev, Wang Ling, Sona Mokra, Agustin Dal Lago, Yotam Doron, Susannah Young, Phil Blunsom, and Chris Dyer. 2020. The DeepMind Chinese–English document translation system at WMT2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 326–337, Online. Association for Computational Linguistics.

Jingyi Zhang and Josef van Genabith. 2020. Translation quality estimation by jointly learning to score and rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2592–2598, Online. Association for Computational Linguistics.

A Ablation Study

Variant	Segment-level Kendall's tau	Pairwise accuracy	Magnitude of scores	BLEU (newstest21) EN−DE	BLEU (newstest21) ZH−EN
Prism (no fine-tuning)	23.5	78.7	0.47	25.6	18.7
Prism+FT	26.8	76.7	0.37	23.0	21.0
– without cross-entropy	26.6	74.6	0.35	10.2	9.6
– without forward ranking	26.0	79.2	0.40	21.9	20.1
– without backward ranking	25.6	77.7	0.39	21.1	20.3

Table A1: Ablation study for the proposed fine-tuning objective, based on the in-domain meta-evaluation setting (WMT 2021 news translations). In every row we remove one aspect of the fine-tuning setup. Meta-metrics are averaged across three language pairs. *Magnitude of scores* refers to the average segment-level scores predicted by the Prism model, converted to probability space via 2^x.

B Evaluation on TED Talks

	EN−DE		EN−RU		ZH−EN		Average	
	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}
BLEU	13.4	7.1 **(-47.0%)**	16.0	12.8 **(-20.0%)**	11.0	9.1 **(-17.3%)**	13.5	9.7 **(-28.2%)**
chrF	14.3	7.9 **(-44.8%)**	18.9	12.8 **(-32.3%)**	11.4	9.0 **(-21.1%)**	14.9	9.9 **(-33.4%)**
Prism	23.6	17.7 **(-25.0%)**	22.0	17.5 **(-20.5%)**	18.0	15.9 **(-11.7%)**	21.2	17.0 **(-19.7%)**
Prism+FT	26.4	24.2 **(-8.3%)**	22.2	21.6 **(-2.7%)**	20.2	19.4 **(-4.0%)**	22.9	21.7 **(-5.2%)**
COMET	27.3	24.6 **(-9.9%)**	25.8	23.2 **(-10.1%)**	20.8	20.7 **(-0.5%)**	24.6	22.8 **(-7.3%)**

Table A2: Segment-level correlation of MT metrics when provided with the standard references and with machine-translated references. The percentages denote the relative change in correlation when falling back to machine-translated references.

	EN−DE		EN−RU		ZH−EN		Average	
	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}
BLEU	66.7	35.9 **(-46.2%)**	83.5	58.2 **(-30.3%)**	64.1	65.4 **(2.0%)**	71.4	53.2 **(-25.6%)**
chrF	65.4	46.2 **(-29.4%)**	85.7	53.8 **(-37.2%)**	61.5	66.7 **(8.5%)**	70.9	55.6 **(-21.6%)**
Prism	69.2	44.9 **(-35.1%)**	82.4	48.4 **(-41.3%)**	67.9	66.7 **(-1.8%)**	73.2	53.3 **(-27.1%)**
Prism+FT	66.7	51.3 **(-23.1%)**	81.3	61.5 **(-24.4%)**	62.8	70.5 **(12.3%)**	70.3	61.1 **(-13.0%)**
COMET	84.6	53.8 **(-36.4%)**	78.0	74.7 **(-4.2%)**	67.9	75.6 **(11.3%)**	76.8	68.0 **(-11.5%)**

Table A3: System-level pairwise accuracy of MT metrics when provided with the standard references and with machine-translated references.

C Training Data Statistics

Language pair	**EN–DE**	**ZH–EN**
Number of systems (including sets of human translations)	10	10
Number of annotated segments	1 418	2 000
– used for relative rankings	1 411	1 985
Number of annotated system translations	14 110	19 994
– used for relative rankings	14 110	19 850
Number of relative rankings	126 217	164 137
– training split	121 217	159 137
– validation split	5 000	5 000

Table A4: Statistics for the WMT 2020 MQM ratings (Freitag et al., 2021a) and for the relative rankings that we derive using an intra-annotator pairing approach.

D Meta-Evaluation Data Statistics

	News			TED Talks		
	EN–DE	EN–RU	ZH–EN	EN–DE	EN–RU	ZH–EN
Number of systems (without human)	13	14	13	13	14	13
Number of MQM-annotated segments	527	527	650	529	512	529
Number of segments with machine-translated reference (on average across system pairs)	518	527	461	517	511	505

Table A5: Statistics for the WMT 2021 MQM ratings (Freitag et al., 2021b) we use for evaluating the metrics.

E Model Hyperparameters

Model	N	d_{model}	d_{ffn}	h	Parameters	Vocabulary size
Prism (Thompson and Post, 2020)	16	1280	12288	20	745M	64k
`wmt21-comet-mqm` (Rei et al., 2021)	24	1024	4096	16	581M	250k

Table A6: Hyperparameters of the Transformer-based metrics.

F Additional Examples of Human-created and Machine-translated References

English–German News Example

Source sequence:
Face masks are mandatory across the state of California, even in fresh air.

Standard reference:
Gesichtsmasken sind im ganzen Bundesstaat Kalifornien vorgeschrieben, auch im Freien.

Randomly sampled error-free system translation (Nemo):
Gesichtsmasken sind im gesamten Bundesstaat Kalifornien Pflicht, auch an der frischen Luft.

Chinese–English News Example

Source sequence:
他已承认，是自己在教堂里点火。

Standard reference:
The parish volunteer has admitted that he had started the fire in the church.

Randomly sampled error-free system translation (metricsystem5):
He has admitted that it was himself who set the fire in the church.

English–German TED Talks Example

Source sequence:
Today I'd like to show you the future of the way we make things.

Standard reference:
Ich möchte Ihnen heute zeigen, wie wir in Zukunft Dinge herstellen werden.

Randomly sampled error-free system translation (Online-W):
Heute möchte ich Ihnen die Zukunft der Art und Weise zeigen, wie wir Dinge herstellen.

Chinese–English TED Talks Example

Source sequence:
今天我想向各位展示未来我们制作东西的方式。

Standard reference:
Today I'd like to show you the ways we make things in the future.

Randomly sampled error-free system translation (metricsystem1):
Today I want to show you how we will make things in the future.