Quality and Quantity of Machine Translation References for Automatic Metrics

Vilém Zouhar
ETH Zürich
vzouhar@ethz.ch

Ondřej Bojar
Charles University
bojar@ufal.cuni.cz

Abstract

Automatic machine translation metrics typically rely on *human* translations to determine the quality of *system* translations. Common wisdom in the field dictates that the human references should be of very high quality. However, there are no cost-benefit analyses that could be used to guide practitioners who plan to collect references for machine translation evaluation. We find that higher-quality references lead to better metric correlations with humans at the segment-level. Having up to 7 references per segment and taking their average (or maximum) helps all metrics. Interestingly, the references from vendors of different qualities can be mixed together and improve metric success. Higher quality references, however, cost more to create and we frame this as an optimization problem: given a specific budget, what references should be collected to maximize metric success. These findings can be used by evaluators of shared tasks when references need to be created under a certain budget.

1. Introduction

Machine translation systems are robustly evaluated through human annotation. This is non-scaleable and non-replicable (Freitag et al., 2021a) for settings such as shared tasks where a number of teams submit automatic translations of the same testset. Automatic metrics aim to provide a cheap and replicable solution. Given the translation and possibly the source and reference segments, they produce a score that correlates with what a human annotator would predict. There is support and evidence for not using references (Lommel, 2016) in metrics, i.e. quality estimation (Specia and Shah, 2018; Rei et al., 2021). Still, most of the commonly used metrics (Section 3) require human reference translations (Freitag et al., 2023). These metrics work by comparing either the overlap on the surface-level (e.g. BLEU, Papineni et al., 2002), of semantic representations (e.g. COMET, Rei et al., 2020) or some downstream task (e.g. MTEQA, Krubiński et al., 2021).

Humans also do not always arrive at perfect translations and thus the quality of the references themselves varies (Castilho et al., 2018). In cases of very poor translations, such as non-translation,[1] the reference-based metrics would clearly fail. While low-quality references are known to decrease the metric correlations (Freitag et al., 2023), the extent of this effect and interactions with other phenomena remains unclear. Many automatic machine translation metrics support multiple references for a single translation natively or by using an aggregation such as the average. For phrase-based MT and BLEU, the trade-off between the number of references vs. the test set size was studied by Bojar et al. (2013, Section 5), concluding that a single-reference test set of 3000 sentences can be comparable to 6–7 references with just 100–200 test sentences. The usefulness of multiple references was later disputed (Freitag et al., 2020) for state-of-the-art system evaluation and some recent metrics do not even support multiple references. Additionally, a professional experienced translator is likely to produce a better translation than an average crowd-worker. However, the cost of a high-quality human translation is likely also much higher.

In this paper, we aim to quantify the trade-off between reference **quality, quantity** and **cost** for segment-level automatic metric performance. We base our experiments on a small-scale English→Czech dataset with multiple references of varying qualities.

We pose research questions with immediate implications for practitioners. The short answers here are only summaries.

Q: Are higher-quality references useful for automatic evaluation?
A: Low-quality should be avoided. Too much investment has diminishing returns. (Sec. 4.1)
Q: Are multiple references useful?
A: Yes. Averaging or taking the maximum across reference improves the metrics. (Sec. 4.2)
Q: How to allocate the budget?
A: By not focusing exclusively on either quality or quantity of references but their combination. This can be computed by Algorithm 1, given a list of vendors and their attributes. (Sec. 4.4)

2. Related Work

Reference quality is known to affect machine translation evaluation. Freitag et al. (2023) note that very low-quality references reduce metric success.

[0] github.com/ufal/optimal-reference-translations
hf.co/datasets/zouharvi/optimal-reference-translations

[1] Text left untouched in the source language.

Proceedings of the HumEval @LREC-COLING 2024, pages 1–11
Turin, Italy. May 21st, 2024.

This stands in contrast to the pre-neural machine translation era where the reference quality did not play an important role in certain settings (Hamon and Mostefa, 2008). This is likely caused by the much higher quality of systems being compared. Vernikos et al. (2022) hypothesize that ambiguous and vague references are the culprits of metric success deterioration. Additionally, Freitag et al. (2020) study how to avoid low-quality references in human translation campaigns.

The BLEU metric (Papineni et al., 2002) was intended to be used with multiple metrics, which was only rarely put in practice over the years. Nevertheless, newer and more sophisticated methods exist to incorporate them (Qin and Specia, 2015). Our results from Figure 1 confirm the older observations of Finch et al. (2004) or Bojar et al. (2013, Section 5) who study the effect of the reference count on metric performance. Finally, multiple references can be used in training better machine translation systems (Madnani et al., 2008; Zheng et al., 2018; Khayrallah et al., 2020; Mi et al., 2020) or for analyzing model uncertainty (Ott et al., 2018) or evaluation uncertainty (Zhang and Vogel, 2004, 2010; Fomicheva et al., 2020). It is also used outside of machine translation for measuring consensus (Vedantam et al., 2015).

The budget allocation algorithm is reminiscent of active learning or data selection. In machine translation, this is limited to selecting training examples (Haffari et al., 2009; González-Rubio et al., 2012; van der Wees et al., 2017; Shi and Huang, 2020; Mendonça et al., 2023). We focus on algorithmic data selection for higher-quality *evaluation*. We aim to complete similar works on practical advice on machine translation evaluation. Kocmi et al. (2021, 2024) study the reliability of metrics from the perspective of deployment decisions. We show that the configuration of references can make these metrics stronger or weaker on segment-level.

3. Setup

To evaluate the effect of references on automatic machine translation evaluation, we need data with controlled references and reference-based metrics.

Optimal Reference Translations. Zouhar et al. (2023) re-annotate a subset of the English→Czech testset from the News domain of the WMT2020 campaign (Barrault et al., 2020). New references were created by translating the original source in 4 different human settings ranging from generic translation vendors to translatology academics following a novel protocol leading to so-called "optimal reference translations" (Kloudová et al., 2023). This phase was followed by a human annotation and post-editing phase performed by 11 annotators of varying professionalities.

Zouhar et al. (2023) study whether the human quality of the references is really the highest achievable one. They stop short of evaluating the impact of this on machine translation evaluation. We repurpose their data and system submissions from Barrault et al. (2020). We refer to the references, from lowest to highest quality of the source, as R1, R2, R3, and R4. Specifically, R1 to R2 come from standard translation vendors,[2] R3 is high-quality translation vendor, and R4 is the work of translatologists (the optimal reference). See Table 1 for basic statistics.

Source segments & documents	160 & 20
Average source segment length	34 tokens
Reference segments	$160 \times 4 = 640$
Reference post-editing	$160 \times 4 \times 11 = 7040$
Systems & system segments	13 & $160 \times 13 = 2080$

Table 1: Overview of the used dataset.

Automated Metrics.[3] For the metrics, we use BLEU (Papineni et al., 2002), chrF (Popović, 2015), TER (Snover et al., 2006), COMET_{20} (Rei et al., 2020), its referenceless version $\text{COMET}_{20}^{\text{QE}}$, and its updated iteration COMET_{22} (Rei et al., 2022), and BLEURT (Sellam et al., 2020). We select these as a representative set of widely-used string-matching and trainable metrics.

Metric Evaluation. We focus on and evaluate metric success at the segment-level ("sentence"-level) by correlating the metric scores with human scores using Kendall's τ.[4] Each translation receives a human score and automatic metric scores which are correlated. This is the standard segment-level evaluation adopted by the WMT Metrics Shared Task (Freitag et al., 2021b, 2022, 2023). In our case (WMT2020), the human segment-level judgments were created from Direct Assessment (Graham et al., 2016) judgements following the "DARR" conversion as described by Mathur et al. (2020): Candidate translations from MT systems were scored on their own, independently of other candidates. For each pair of judgements of candidates translating the same source, we construct one golden-truth item of pairwise comparison if the two individual scores differ by more than 25% absolute. As Mathur et al. (2020), we believe that this difference in the judgement is big enough to trust the simulated pairwise comparison.

[2] Nevertheless, based on observations of Kloudová et al. (2021), R1 are to a large extent post-edits of one of the participating systems.

[3] `BLEU|#:1|c:mixed|e:yes|tok:13a|s:exp`
`chrF|#:1|c:mixed|e:yes|nc:6|nw:0|s:no`
`TER|#:1|c:lc|t:tercom|nr:no|pn:yes|as:no`

[4] $\tau = (\#\text{concordant} - \#\text{discordant})/\#\text{pairs}$; read more on the definition in Macháček and Bojar (2014).

Metric	R1	R2	R3	R4
BLEU	0.082 _	0.103 ▂	0.109 ▂	0.103 ▂
chrF	0.090 _	0.125 ▃	0.128 ▃	0.123 ▃
TER	0.082 _	0.092 _	0.114 ▂	0.105 ▂
COMET_{20}	0.172 ■	0.176 ■	0.185 ■	0.181 ■
COMET_{22}	0.189 ■	0.195 ■	0.191 ■	0.192 ■
BLEURT	0.159 ■	0.156 ■	0.199 ■	0.178 ■
Average	0.129 ▃	0.141 ▃	0.154 ■	0.147 ■
$\text{COMET}_{20}^{\text{QE}}$		0.171 ■		

Table 2: Segment-level Kendall's τ between automatic metrics and human scores. The metrics are computed with respect to each of the four references. The black boxes indicate the value visually and comparable across both columns and rows. The R3 translation yields the best results as the reference, despite not being the optimal translation from the human perspective.

Proficiency	R1^{PE}-R1	R2^{PE}-R2	R3^{PE}-R3	R4^{PE}-R4
Layman	+0.019	+0.011	+0.011	+0.011
Student	+0.009	+0.005	+0.001	−0.002
Professional	+0.025	+0.011	+0.004	+0.002

Table 3: Difference in Kendall's τ between using original translations (in Table 2) and their post-edited versions. The post-editing comes from translators on different levels. The correlations are averaged across all metrics; see Tables 6 and 10 for per-metric breakdowns. In most cases, using post-edited versions improves metric performance.

4. Experiments

4.1. Reference Quality is Important

As stated in Section 3, we have access to four human translations of varying quality. In Table 2, we show the metric success measured by correlation with human scores. The metrics stay the same but the references they use are changed. Across both string-matching and parametrized model-based metrics, R1, the worst human translation, leads to the worst metric performance. The best performance is achieved with R3, a standard professional translation. Notably, it is not R4 which was created by professional translatologists and was also the most expensive one. This can be explained by the presence of translation shifts, which occur more frequently on this professionality level, but can negatively impact the utility of the reference (Fomicheva, 2017). Translation shifts in general refer to deviation from the original structure or meaning. For our new references, the translatologists paid attention to preserve the meaning but they often restructured the sentences. They did this to avoid translationese as much as possible and to express the subtleties of information structure (given vs. new information) which is natively expressed via Czech word order. These boosted word order differences make it harder for automatic metrics to match the, rather translationese, candidate and the reference. We anticipate that more fluent large language model-based MT could sound less translationese and the "optimal reference translations" will serve better in this setting. See Section 5 for an example and analysis.

Aggregation	R3	R{3,4}	Rx	Rx^{PE}
Average	0.154 ■	0.159 ■	0.166 ■	0.164 ■
Max	0.154 ■	0.155 ■	0.165 ■	0.167 ■

Table 4: Average performance of metrics with multiple references. See Table 9 for per-metric breakdown. All aggregation methods improve the performance over the best single one, R3.

A simple way of improving a translation is to post-edit (refine) it, which is cheaper that translating it from scratch (Daems and Macken, 2020; Zouhar et al., 2021). Moreover, Bojar et al. (2013, Figure 7) show that such post-edited references lead to a better performance of BLEU, because "every n-gram mismatch indicates an error". With standard references, an n-gram mismatch often means just lack of reference coverage. However, such post-edited references need to be ideally created for each evaluated MT system. In our case, the post-edits were created starting from *human* reference translations Rx. We mark them Rx^{PE} and use them as references for the automatic metrics in Table 3. The post-editors are either laymen with knowledge of both languages, students of translatology, or professional translators. While the proficiency level plays a role, in most cases the post-edited translations serve as better references. Table 6 below lists the raw metric score changes in a closer detail.

4.2. Multiple References are Useful

The previous section provided an analysis of how individual references affect metric performance. In many situations, however, multiple references are available. While some metrics, such as BLEU, support multiple references natively, one can also aggregate them using either segment-level averages or maxima (i.e. compute multiple scores for each segment and take the average or maximum). In Table 4 we consider three setups: (1) two high-quality references, R3 and R4, (2) all human translations, Rx, or (3) all post-edited human translations, Rx^{PE}. Across all metrics, this segment-level aggregation improves the correlation with humans, especially in the case of using the original four human translations. Taking the maximum and not the average has the advantage that there exists a specific reference which yields that particular score. The

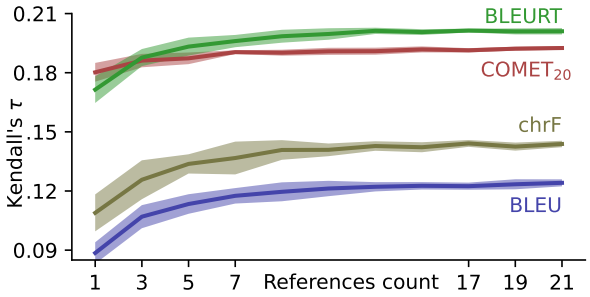

Figure 1: Metric performance with multiple sampled references from the pool of the original human translations and their post-edited versions. Confidence t-test intervals indicate 99% confidence of the mean (of 10 samples) being in the shaded area. 👁 Biggest advantage is gained from at least three references and taking their segment-level average (max aggregation not shown).

maximum also reflects the spirit of automated evaluation: measuring some similarity between the candidate and reference translations. With more references, taking the maximum corresponds to first finding the most similar reference. We include all subsets as references in Table 9.

To systematically study the effect of reference count on metric performance, in Figure 1 we randomly sample x references from the whole pool of original and post-edited translations, irrespective of their quality. The biggest gains in metric performance are achieved until seven references and further gains are negligible, which is in line with the observations of Bojar et al. (2013, Section 5).

Metric	R1	R2	R4	R3	R1^{PE}	R3^{PE}
BLEU	24.2	31.5	27.3	37.1	23.9	31.0
chrF	55.7	60.3	56.1	63.0	54.5	58.4
TER	-63.3	-53.0	-59.4	-48.7	-64.1	-58.9
COMET^{20}	65.5	68.9	61.0	68.2	60.4	61.4
COMET^{22}	84.6	84.9	83.6	84.8	83.6	83.7
BLEURT	61.3	66.1	64.5	68.8	61.6	64.9

Table 5: Raw average scores across metrics and references. TER scores are flipped to make higher numbers be better. The columns are sorted by quality of references from worst to best as reported in Table 2. 👁 For most metrics, higher absolute metric scores correspond to better evaluation (numbers are growing from left to right), except for post-edited human references Rx^{PE} which serve better as references (are more to the right) but lead to lower absolute metric scores.

4.3. References and Metric Scores

To understand the effect of different metrics, we show the average *raw* scores of each metric in Table 5. While it appears that the higher the raw score, the better the metric performance (low score of R1 and high scores of R3 and R4), this trend does not explain the improvements of using the post-edited versions, e.g. as R1^{PE} over R1, or R3^{PE} over R3. In fact, the post-edited versions always lead to lower raw scores. This could be the result of either further translation shifts as the post-edits are based on a translation and not the source or additional (fully justified) corrections in the references which lead to fewer matches with the candidates.

4.4. Allocating a Budget for References

Usually, it is simple to gather many source sentences and let multiple systems translate them. Evaluating all of them using human annotators is unattainable but running automatic metrics is not. However, these require references, which are also costly. It remains unclear how many references and of which quality to obtain to achieve the most reliable automatic quality assessment under a given budget.

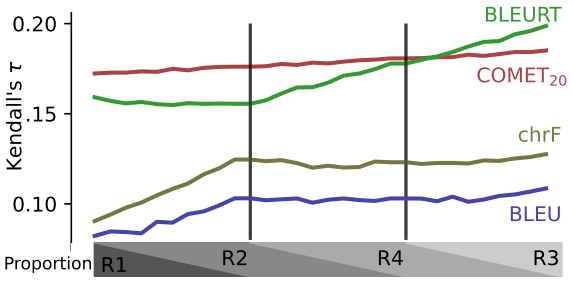

Figure 2: Metric performance with references (ordered by usefulness) from mixed sources (e.g. 25% R1 and 75% R2; rightmost is 100% R3). 👁 Mixing references does not hurt any metric.

Can references be mixed? To assess what types of configurations of references *can* lead to the most reliable automatic evaluation, we first validate if references can be meaningfully mixed. For example, if it is viable that 75% of the sources can have references from a cheaper vendor R1 and 25% from a higher-quality but more expensive vendor R3. This is different from Table 4 where each segment had exactly two references from the same two sources. In Section 4.3, we show that using lower-quality references R1 leads to lower absolute metric scores (e.g. BLEU = 24.2) as opposed to higher-quality ones R3 (e.g. BLEU = 37.1). This holds across all metrics. Bojar et al. (2010) observe that lower BLEU scores are less reliable, but they refer to the range of BLEU < 20. It is thus questionable if BLEUs at 20–40 correlate differently with human MT quality judgements. In Figure 2, we mix some of the references together for evaluation, but staying at single-reference evaluation.

Algorithm 1 Budget Allocation for References

Input: Source segments S, levels L, cost function $\textsc{Cost} : L \to \mathbb{R}^+$, utility function $\textsc{Util.} : L \to \mathbb{R}^+$, tradeoff hyperparameter $\lambda \in [0, 1]$, temperature $t > 0$, budget $B \in \mathbb{R}^+$.

Output: Assignment $R : L \to 2^S$.

Note: Figure 4 contains a patience mechanism instead of exit on error.

$L \leftarrow \textsc{Sort}(L, \textsc{Cost})$

$R[L_0] \leftarrow S$; $O \leftarrow R$ ▷ Assign everything to the cheapest level at first.

while $\sum_{l \in L} |R[l]| \cdot \textsc{Cost}(l) < B \ \wedge$ **no exception do**

  $O \leftarrow R$

  $a \sim \textsc{Sample}(\textsc{Promote} : \lambda, \textsc{Add} : 1 - \lambda)$ ▷ Select action.

  $X^+ \leftarrow \{\langle s, l\rangle | l \in L, s \in S \setminus R[l]\}$ ▷ Samples that could be added to $R[l]$.

  $X^- \leftarrow \{\langle s, l\rangle | l \in L, s \in R[l]\}$ ▷ Samples that could be removed from $R[l]$.

  if $a = \textsc{Add}$ **then**

    $x, l \sim \textsc{Sample}(\{\langle x, l\rangle : \frac{\sigma(\textsc{Util.}(l) - \textsc{Cost}(l))^{1/t}}{Z} | x, l \in X^+\})$ ▷ Sample a segment to add.

    $R[l] \leftarrow R[l] \cup \{x\}$ ▷ Commit transaction.

  else if $a = \textsc{Promote}$ **then**

    $x^+, l^+ \sim \textsc{Sample}(\{\langle x, l\rangle : \frac{\sigma(\textsc{Util.}(l) - \textsc{Cost}(l))^{1/t}}{Z} | x, l \in X^+\})$ ▷ Sample a segment to add.

    ▷ Sample where to move from.

    $\_ , l^- \sim \textsc{Samp.}(\{\langle x, l\rangle : \frac{\sigma(\textsc{Cost}(l) - \textsc{Util.}(l))^{1/t}}{Z} | x, l \in X^-, x{=}x^+ \wedge \textsc{Util.}(l^-){<}\textsc{Util.}(l^+)\})$

    $R[l^+] \leftarrow R[l^+] \cup \{x^+\}$; $R[l^-] \leftarrow R[l^-] \setminus \{x^+\}$ ▷ Commit transaction.

  end if

end while; **return** O

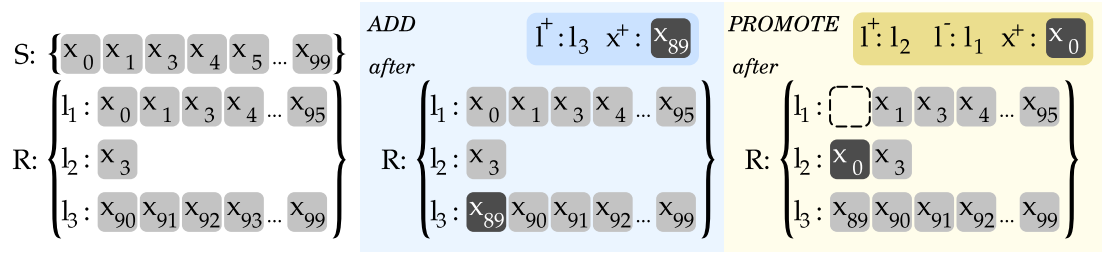

Figure 3: Illustration of two operations from Algorithm 1. The initial state is on the left. Then, a new segment x_{89} is added to the l_3 level. Lastly, the segment x_0 is promoted from l_1 to l_2.

Despite the varying absolute scores of metrics under different references, as explored in Section 4.3, mixing of multiple references leads to an almost perfectly linear combination of the endpoint metric performances. The biggest gains in this respect are obtained by BLEURT, chrF and BLEU, while COMET_{20} is almost unaffected. There is no formal guarantee that the mix of score distributions will not lower the overall Kendall's τ. Nevertheless, a positive conclusion is that if there is budget to only translate 25% of segments with high quality, it should be done and it can only improve the overall evaluation reliability.

Budget Allocation Algorithm. We provide a heuristic stochastic algorithm to find an assignment of source segments S to be translated by vendors of different costs and qualities within a specific budget. For the current dataset, we set the cost of a segment in R1, R2, R3, and R4 to be 1, 1, 2, and 3, respectively. Their quality (or "fitness" for the purpose of automatic evaluation) were set to 1, 2, 4, and 3 based on our observations in Section 4.1. Algorithm 1 contains a hyperparameter λ that controls whether the budget will be allocated more towards having multiple references per-segment or more towards having fewer but higher-quality references per-segment, and the temperature t than controls the the sampling randomness.

We formalize the problem with a segment cost $\textsc{Cost}(l)$ for a reference on level $l \in L$ and the utility $\textsc{Util.}(l)$. The levels might correspond to translation vendors which have costs and qualities. In our case, $\textsc{Cost} = \{\text{R1}{:}\,1, \text{R2}{:}\,1, \text{R3}{:}\,2, \text{R4}{:}\,3\}$ and $\textsc{Util.} = \{\text{R1}{:}\,1, \text{R2}{:}\,2, \text{R3}{:}\,4, \text{R4}{:}\,3\}$. Given a set of source sentences S, the goal is to assign the segments to different levels $\text{R1} \ldots \text{R4}$. The same segment can be assigned to different quality levels at once, leading to multiple references for that segment. The selection should maximize performance of a particular metric on a number of systems but needs to fit under a fixed budget B, i.e.

$\sum_l |R_l| \cdot \text{COST}(l) \leq B$. In our setup, to preserve fair comparison, each segment needs to have at least one reference. This is because the smaller the testset, the easier it is to achieve higher but spurious correlations. Therefore, $\bigcup_{l \in L} R_l = S$. The formalization explicitly allows for parts of the testset to be translated multiple times but requires the budget to cover at least the full test set with the cheapest references. This requirement can be fulfilled by subsampling the testset, as commonly done in WMT evaluation campaigns (Kocmi et al., 2023, inter alia).

The pseudocode is provided in Algorithm 1 and explanatory illustration of the two operations in Figure 3. The algorithm continually applies one of the two operations until they can either no longer be applied or the budget is reached. The algorithm will always terminate because ADD increases the cost and utility and PROMOTE increases the utility by at least $\min_{l \in L} \text{COST}(l)$ and $\min_{l_1, l_2 \in L} |\text{UTIL.}(l_2) - \text{UTIL.}(l_1)|$, respectively. Therefore either the budget will be filled or every segment will receive a reference from all vendors.

In Figure 4, we show chrF and COMET_{20} correlations when using the references selected by our algorithm. The optimal preference between quality and quantity changes with increasing budget. Using all of the budget on either quality or quantity would correspond to the bottom or top row, which are not optimal. The best reference configurations for a particular budget, such as $|S| \times 4$, four-times the price of the cheapest translation, contain a mixture of references from R1, R2, R3, and R4 with multiple references for some segments. In addition to the metric correlations in Figure 4, we show the average number of references per source segment in Figure 5. With focus on quality, each segment has fewer references.

5. Qualitative Analysis

In Tables 7 and 8, we show a single source segment, one system translation and multiple references and the metric scores. BLEU ranges from 0 to 100 and both extremes are almost achieved just with a different reference. The best human translation led to the lowest BLEU score because of a translation shift. This is not surprising because BLEU operates on the surface-level. Unexpectedly, a similar thing happens also with COMET_{20}, which uses a distributed semantic representation of the segments. This shows that parametric model-based metrics are not robust to changes in references. In Table 8, the COMET_{20} difference between references is large due to some translators deciding to drop the verb *"spolupracovat" (collaborate)*, which changes the meaning and the system translation is penalized.

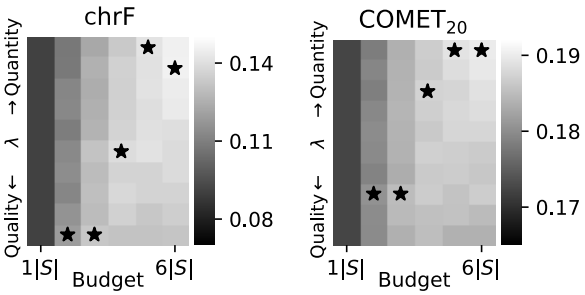

Figure 4: Heatmaps of chrF (left) and COMET_{20} (right) Kendall's τ correlations on reference configurations created with a specific budget (x-axis) and quality-quantity trade-off λ (y-axis). ★ marks the best value in each column (fixed budget). The first column corresponds to the cheapest translation for all test segments, with no room for selection. $\lambda \in [0, 0.7]$ and $t = 0.5$. With a limited budget, e.g. $2|S|$ or $3|S|$, it makes more sense to add *some* references of a higher quality rather than covering the whole test set with a second reference. With more budget available, multiple references per segment become more beneficial.

6. Conclusion

We showed that the quality of references is important for accurate automatic machine translation metrics. The relationship is not straightforward: translatologists' translations, despite being the peak translation quality, are not the best references. Rather, it is the *standard commercial professional translations* that work best for current metrics. The trend applies to both string-matching metrics as well as to parametric model-based ones. Taking the *average over multiple references provides the biggest benefit*, with diminishing returns after 7 references. We also provided a heuristic-based *algorithm for finding a good configuration of references given a budget*, which surpasses optimizing solely for quantity or quantity.

Future work. The dataset size prevents system-level investigations. Because there is little point in evaluating segments that are easy to translate, a follow-up approach could prioritize difficult-to translate segments. This is used by Isabelle et al. (2017) for creating a challenge set. Future works should *quantify* the references quality and ask how many segments are needed to fulfill a certain desideratum, such as effect size or metric accuracy.

Limitations. We note the limitation of using a small dataset and a single language translation direction due to the costs of creating multiple rounds of high-quality references. We are convinced the results hold in other scenarios as the effect directions are the same across multiple metrics and setups.

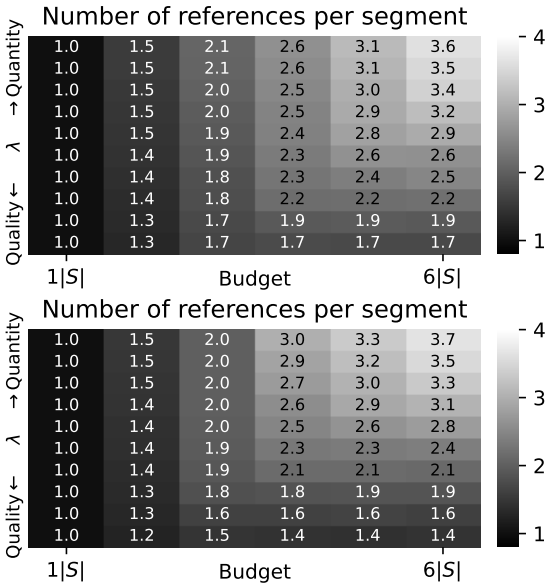

Figure 5: Average number of references per one segment allocated by Algorithm 1 with $\tau = 0.5$ (top) and $\tau = 10^{-3}$ (bottom).

	Metric	R1^{PE}-R1	R2^{PE}-R2	R3^{PE}-R3	R4^{PE}-R4
Layman PE	BLEU	+0.019	+0.007	+0.009	+0.010
	chrF	+0.027	+0.015	+0.016	+0.019
	TER	+0.026	+0.015	+0.013	+0.014
	COMET^{20}	+0.016	+0.011	+0.010	+0.009
	COMET^{22}	+0.008	+0.006	+0.007	+0.005
	BLEURT	+0.017	+0.015	+0.011	+0.010
Student PE	BLEU	+0.010	+0.001	−0.004	−0.001
	chrF	+0.011	+0.001	+0.000	−0.004
	TER	+0.003	+0.001	+0.005	−0.002
	COMET^{20}	+0.010	+0.003	−0.002	−0.002
	COMET^{22}	+0.002	+0.001	+0.000	−0.002
	BLEURT	+0.021	+0.022	+0.005	−0.004
Prof. PE	BLEU	+0.035	+0.011	+0.007	−0.000
	chrF	+0.040	+0.010	+0.008	+0.004
	TER	+0.023	+0.014	+0.003	−0.002
	COMET^{20}	+0.016	+0.006	+0.000	+0.005
	COMET^{22}	+0.008	+0.002	+0.003	+0.005
	BLEURT	+0.027	+0.022	+0.005	−0.001

Table 6: Difference between using original translations (in Table 2) and post-edited translations as references. Sections are divided based on who did the post-editing (layman, translatology student, or professional translator). This table expands on Table 3. Absolute scores of individual reference subsets are in Table 10.

Table 7: BLEU and COMET$_{20}$ scores of the source "*Three Scottish students named among Europe's best*" and the system translation "*Tři skotští studenti byli zařazeni mezi nejlepší v Evropě*". Both metrics are multiplied by 100. 👁 All references are good translations but the scores vary.

BLEU	COMET$_{20}$	Reference
10	78	K evropské špičce nově patří i tři skotští studenti
23	120	Tři skotští studenti se umístili mezi nejlepšími v Evropě
23	121	Tři skotští studenti mezi nejlepšími v Evropě
28	116	Tři skotští studenti byli oceněni jako jedni z nejlepších v Evropě
28	115	Tři skotští studenti byli jmenováni jako jedni z nejlepších v Evropě
28	114	Tři skotští studenti byli vyhlášeni jako jedni z nejlepších v Evropě
32	117	Tři skotští studenti byli jmenováni jednimi z nejlepších v Evropě
37	122	Tři skotští studenti byli jmenováni mezi nejlepšími v Evropě
43	125	Tři skotští studenti se zařadili mezi nejlepší v Evropě
43	121	Tři skotští studenti patří mezi nejlepší v Evropě.
43	122	Tři skotští studenti patří mezi nejlepší v Evropě
60	127	Tři skotští studenti zařazeni mezi nejlepší v Evropě
100	131	Tři skotští studenti byli zařazeni mezi nejlepší v Evropě

Table 8: BLEU and COMET$_{20}$ scores of the source "*Sony, Disney Back To Work On Third Spider-Man Film*" and the system translation "*Disney se vrací, bude spolupracovat se Sony na třetím sólovém Spider-Man filmu*". Both metrics are multiplied by 100. 👁 Some references omit part of the information and COMET$_{20}$ thus penalizes the system translation.

BLEU	COMET$_{20}$	Reference
4	-42	Sony a Disney točí třetí film o Spidermanovi
4	-33	Sony a Disney točí třetí film o Spider-Manovi
8	-9	Sony a Disney pracují na třetím filmu o Spider-Manovi
8	-4	Sony a Disney pokračují v práci na třetím filmu o Spider-Manovi
8	1	Sony a Disney opět pracují na třetím filmu o Spider-Manovi
8	15	Sony a Disney spolupracují na třetím filmu o Spider-Manovi
4	16	Sony a Disney budou spolupracovat při natáčení třetího filmo o Spider-manovi
8	28	Sony a Disney opět spolupracují na třetím filmu o Spider-Manovi
8	30	Sony a Disney budou spolupracovat na třetím filmu o Spider-Manovi
8	35	Sony a Disney budou opět spolupracovat na třetím filmu o Spider-Manovi
17	52	Disney bude znovu spolupracovat se společností Sony na třetím filmu Spider-Man
10	64	Disney bude se Sony dál pracovat na třetím filmu se Spider-Manem
50	73	Disney bude spolupracovat se Sony na třetím sólovém filmu o Spider-Manovi
75	99	Disney se vrací, bude spolupracovat se Sony na třetím filmu o Spider-Manovi
78	106	Disney se vrací, bude spolupracovat se Sony na třetím sólovém filu Spider-Man.
79	108	Disney se vrací, bude spolupracovat se Sony na třetím Spider-Man filmu
100	121	Disney se vrací, bude spolupracovat se Sony na třetím sólovém Spider-Man filmu

		R1	R2	R3	R4	R{1,2}	R{1,3}	R{1,4}	R{2,3}	R{2,4}	R{3,4}	R{1,2,3}	R{1,2,4}	R{1,3,4}	R{2,3,4}	Rx
Average	BLEU	0.082	0.103	0.109	0.103	0.109	0.122	0.114	0.132	0.124	0.114	0.136	0.124	0.125	0.130	0.134
	chrF	0.090	0.125	0.128	0.123	0.121	0.135	0.124	0.148	0.139	0.135	0.146	0.135	0.140	0.147	0.147
	TER	0.082	0.092	0.114	0.105	0.095	0.120	0.107	0.125	0.117	0.120	0.121	0.110	0.123	0.127	0.124
	$COMET_{20}$	0.172	0.176	0.185	0.181	0.181	0.189	0.185	0.191	0.183	0.188	0.190	0.185	0.190	0.189	0.189
	$COMET_{22}$	0.189	0.195	0.191	0.192	0.195	0.197	0.194	0.201	0.197	0.195	0.200	0.197	0.197	0.199	0.199
	BLEURT	0.159	0.156	0.199	0.178	0.171	0.203	0.183	0.201	0.180	0.203	0.201	0.184	0.204	0.202	0.202
	Average	0.129	0.141	0.154	0.147	0.145	0.161	0.151	0.166	0.157	0.159	0.166	0.156	0.163	0.166	0.166
Max	BLEU	0.082	0.103	0.109	0.103	0.116	0.122	0.118	0.135	0.132	0.111	0.138	0.137	0.121	0.135	0.137
	chrF	0.090	0.125	0.128	0.123	0.133	0.137	0.129	0.139	0.146	0.135	0.144	0.148	0.140	0.144	0.147
	TER	0.082	0.092	0.114	0.105	0.101	0.124	0.116	0.132	0.128	0.117	0.132	0.130	0.126	0.135	0.134
	$COMET_{20}$	0.172	0.176	0.185	0.181	0.177	0.184	0.185	0.180	0.183	0.183	0.181	0.183	0.184	0.181	0.182
	$COMET_{22}$	0.189	0.195	0.191	0.192	0.195	0.191	0.196	0.191	0.195	0.191	0.192	0.197	0.191	0.192	0.192
	BLEURT	0.159	0.156	0.199	0.178	0.180	0.199	0.190	0.188	0.181	0.193	0.197	0.192	0.200	0.188	0.197
	Average	0.129	0.141	0.154	0.147	0.150	0.159	0.156	0.161	0.161	0.155	0.164	0.164	0.160	0.162	0.165

Table 9: Comparison using either a single or multiple references and taking the average or maximum on segment-level. This table expands on Table 4. The black boxes indicate the reported value of Kendall's τ visually and are comparable across columns as well as rows.

		$R1^{PE}$	$R2^{PE}$	$R3^{PE}$	$R4^{PE}$	Rx^{PE}	$R\{1,1^{PE}\}$	$R\{2,2^{PE}\}$	$R\{3,3^{PE}\}$	$R\{4,4^{PE}\}$	$R\{x,x^{PE}\}$
Layman PE	BLEU	0.101	0.111	0.117	0.113	0.144	0.092	0.107	0.113	0.108	0.140
	chrF	0.118	0.139	0.144	0.142	0.164	0.106	0.134	0.137	0.135	0.159
	TER	0.107	0.107	0.127	0.119	0.142	0.099	0.102	0.123	0.116	0.139
	$COMET^{20}$	0.188	0.187	0.195	0.190	0.198	0.183	0.184	0.193	0.188	0.197
	$COMET^{22}$	0.197	0.201	0.198	0.197	0.203	0.195	0.200	0.196	0.196	0.202
	BLEURT	0.176	0.170	0.210	0.188	0.209	0.169	0.165	0.206	0.186	0.209
	Average	0.148	0.153	0.165	0.158	0.177	0.141	0.149	0.161	0.155	0.174
Student PE	BLEU	0.092	0.104	0.105	0.102	0.123	0.089	0.107	0.108	0.103	0.130
	chrF	0.102	0.126	0.128	0.119	0.139	0.097	0.127	0.130	0.122	0.144
	TER	0.085	0.093	0.119	0.103	0.119	0.084	0.095	0.119	0.104	0.123
	$COMET^{20}$	0.182	0.179	0.183	0.179	0.186	0.179	0.179	0.186	0.181	0.188
	$COMET^{22}$	0.191	0.196	0.191	0.189	0.195	0.191	0.197	0.193	0.191	0.197
	BLEURT	0.180	0.178	0.203	0.174	0.199	0.172	0.170	0.204	0.177	0.202
	Average	0.139	0.146	0.155	0.144	0.160	0.136	0.146	0.157	0.146	0.164
Professional PE	BLEU	0.118	0.114	0.115	0.103	0.127	0.103	0.113	0.113	0.104	0.133
	chrF	0.131	0.135	0.136	0.127	0.146	0.113	0.133	0.135	0.126	0.149
	TER	0.105	0.106	0.116	0.103	0.118	0.094	0.102	0.118	0.104	0.122
	$COMET^{20}$	0.188	0.182	0.185	0.186	0.190	0.183	0.181	0.189	0.185	0.191
	$COMET^{22}$	0.198	0.198	0.195	0.196	0.199	0.195	0.199	0.195	0.195	0.200
	BLEURT	0.186	0.178	0.204	0.176	0.197	0.177	0.172	0.206	0.179	0.202
	Average	0.154	0.152	0.159	0.149	0.163	0.144	0.150	0.159	0.149	0.166

Table 10: Metric performance when using post-edited references also jointly with their original versions (averaged at the segment-level). This table expands on Tables 3 and 6.

Acknowledgements

We extend our gratitude to Yasmin Moslem, Isabella Lai, Theia Vogel, Blanka Sokołowska, Raj Dabre, Tom Kocmi, and Farhan Samir, who proofread this paper in their free time. Ondřej Bojar received funding from Ministry of Education, Youth and Sports of the Czech Republic LM2018101 LINDAT/CLARIAH-CZ and the 19-26934X grant of the Czech Science Foundation (NEUREM3).

Bibliographical References

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 conference on machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, 1–55. Association for Computational Linguistics.

Ondřej Bojar, Kamil Kos, and David Mareček. 2010. Tackling sparse data issue in machine translation evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, 86–91, Uppsala, Sweden. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. *Translation quality assessment: From principles to practice*.

Joke Daems and Lieve Macken. 2020. Post-editing human translations and revising machine translations: Impact on efficiency and quality. In *Translation Revision and Post-editing*, 50–70. Routledge.

Andrew M Finch, Yasuhiro Akiba, and Eiichiro Sumita. 2004. How does automatic machine translation evaluation correlate with human scoring as the number of reference translations increases? In *LREC*.

Marina Fomicheva. 2017. *The role of human reference translation in machine translation evaluation*. Ph.D. thesis, Universitat Pompeu Fabra.

Marina Fomicheva, Lucia Specia, and Francisco Guzmán. 2020. Multi-hypothesis machine translation evaluation. In *Proceedings of the 58th Annual Meeting of ACL*.

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 Conference on Empirical Methods in Natural Language Processing*, 61–71. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, 578–628. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. 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*, 46–68. 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*, 733–774. Association for Computational Linguistics.

Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 245–254. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 415–423. Association for Computational Linguistics.

Olivier Hamon and Djamel Mostefa. 2008. The impact of reference quality on automatic MT evaluation. In *Coling 2008: Companion volume: Posters*, 39–42.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2486–2496. Association for Computational Linguistics.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 82–89. Association for Computational Linguistics.

Věra Kloudová, David Mraček, Ondřej Bojar, and Martin Popel. 2023. Možnosti a meze tvorby tzv. optimálních referenčních překladů: po stopách „překladatelštiny“ v profesionálních překladech zpravodajských textů. *Slovo a slovesnost*, 84(2):122–156.

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)*, 114–119, Stroudsburg, USA. 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*, 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. In *Proceedings of the Sixth Conference on Machine Translation*, 478–494. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021. Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, 495–506, Online. Association for Computational Linguistics.

Arle Lommel. 2016. Blues for BLEU: Reconsidering the validity of reference-based MT evaluation. *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, 63.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 293–301, Baltimore, MD, USA. Association for Computational Linguistics.

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz. 2008. Are multiple reference translations necessary? investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas*, 143–152.

Nitika Mathur, Johnny Tian-Zheng Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Fifth Conference on Machine Translation - Proceedings of the Conference*, 688–725, Stroudsburg, PA, USA. Association for Computational Linguistics, Association for Computational Linguistics.

Vânia Mendonça, Ricardo Rei, Luísa Coheur, and Alberto Sardinha. 2023. Onception: Active Learning with Expert Advice for Real World Machine Translation. *Computational Linguistics*, 49(2):325–372.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2020. Improving adversarial neural machine translation for morphologically rich language. *IEEE Transactions on Emerging Topics in Computational Intelligence*, (4):417–426.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 3956–3965. PMLR.

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 of ACL*, 311–318. 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*, 392–395. Association for Computational Linguistics.

Ying Qin and Lucia Specia. 2015. Truly exploring multiple references for machine translation evaluation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, 578–585. Association for Computational Linguistics.

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*, 1030–1040. 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 Conference on Empirical Methods in Natural Language Processing*, 2685–2702. Association for Computational Linguistics.

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

Zhouxing Shi and Minlie Huang. 2020. Robustness to modification with shared words in paraphrase identification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 164–171. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231. Association for Machine Translation in the Americas.

Lucia Specia and Kashif Shah. 2018. Machine translation quality estimation: Applications and future perspectives. *Translation quality assessment: from principles to practice*, 201–235.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1400–1410. Association for Computational Linguistics.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 118–128. Association for Computational Linguistics.

Ying Zhang and Stephan Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of the 10th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*.

Ying Zhang and Stephan Vogel. 2010. Significance tests of automatic machine translation evaluation metrics. *Machine Translation*, 24:51–65.

Renjie Zheng, Mingbo Ma, and Liang Huang. 2018. Multi-reference training with pseudo-references for neural translation and text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 3188–3197. Association for Computational Linguistics.

Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021. Neural machine translation quality and post-editing performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10204–10214. Association for Computational Linguistics.

Vilém Zouhar, Věra Kloudová, Martin Popel, and Ondřej Bojar. 2023. Evaluating optimal reference translations.