BLEU might be Guilty but References are not Innocent

Markus Freitag, David Grangier, Isaac Caswell
Google Research
{freitag,grangier,icaswell}@google.com

Abstract
The quality of automatic metrics for machine translation has been increasingly called into question, especially for high-quality systems. This paper demonstrates that, while choice of metric is important, the nature of the references is also critical. We study different methods to collect references and compare their value in automated evaluation by reporting correlation with human evaluation for a variety of systems and metrics. Motivated by the finding that typical references exhibit poor diversity, concentrating around translationese language, we develop a paraphrasing task for linguists to perform on existing reference translations, which counteracts this bias. Our method yields higher correlation with human judgment not only for the submissions of WMT 2019 English→German, but also for Back-translation and APE augmented MT output, which have been shown to have low correlation with automatic metrics using standard references. We demonstrate that our methodology improves correlation with all modern evaluation metrics we look at, including embedding-based methods. To complete this picture, we reveal that multi-reference BLEU does not improve the correlation for high quality output, and present an alternative multi-reference formulation that is more effective.

1 Introduction
Machine Translation (MT) quality has greatly improved in recent years (Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017). This progress has cast doubt on the reliability of automated metrics, especially in the high accuracy regime. For instance, the WMT English→German evaluation in the last two years had a different top system when looking at automated or human evaluation (Bojar et al., 2018; Barrault et al., 2019). Such discrepancies have also been observed in the past, especially when comparing rule-based and statistical systems (Bojar et al., 2016b; Koehn and Monz, 2006; Callison-Burch et al., 2006).

Automated evaluations are however of crucial importance, especially for system development. Most decisions for architecture selection, hyperparameter search and data filtering rely on automated evaluation at a pace and scale that would not be sustainable with human evaluations. Automated evaluation (Koehn, 2010; Papineni et al., 2002) typically relies on two crucial ingredients: a metric and a reference translation. Metrics generally measure the quality of a translation by assessing the overlap between the system output and the reference translation. Different overlap metrics have been proposed, aiming to improve correlation between human and automated evaluations. Such metrics range from n-gram matching, e.g. BLEU (Papineni et al., 2002), to accounting for synonyms, e.g. METEOR (Banerjee and Lavie, 2005), to considering distributed word representation, e.g. BERTScore (Zhang et al., 2019). Orthogonal to metric quality (Ma et al., 2019), reference quality is also essential in improving correlation between human and automated evaluation.

This work studies how different reference collection methods impact the reliability of automatic evaluation. It also highlights that the reference sentences typically collected with current (human) translation methodology are biased to assign higher automatic scores to MT output that share a similar style as the reference. Human translators tend to generate translation which exhibit translationese language, i.e. sentences with source artifacts (Koppel and Ordan, 2011). This is problematic because collecting only a single style of references fails to reward systems that might produce alternative but equally accurate translations (Popović, 2019). Because of this lack of diversity, multi-reference evaluations like multi-reference BLEU are also bi-
ased to prefer that specific style of translation.

As a better solution, we show that paraphrasing translations, when done carefully, can improve the quality of automated evaluations more broadly. Paraphrased translations increase diversity and steer evaluation away from rewarding translation artifacts. Experiments with the official submissions of WMT 2019 English→German for a variety of different metrics demonstrate the increased correlation with human judgement. Further, we run additional experiments for MT systems that are known to have low correlation with automatic metrics calculated with standard references. In particular, we investigated MT systems augmented with either back-translation or automatic post-editing (APE). We show that paraphrased references overcome the problems of automatic metrics and generate the same order as human ratings.

Our contributions are four-fold: (i) We collect different types of references on the same test set and show that it is possible to report strong correlation between automated evaluation with human metrics, even for high accuracy systems. (ii) We gather more natural and diverse valid translations by collecting human paraphrases of reference translations. We show that (human) paraphrases correlate well with human judgments when used as reference in automatic evaluations. (iii) We present an alternative multi-reference formulation that is more effective than multi reference BLEU for high quality output. (iv) We release\footnote{https://github.com/google/ wmt19-paraphrased-references} a rich set of diverse references to encourage research in systems producing other types of translations, and reward a wider range of generated language.

2 Related Work

Evaluation of machine translation is of crucial importance for system development and deployment decisions (Moorkens et al., 2018). Human evaluation typically reports adequacy of translations, often complemented with fluency scores (White, 1994; Graham et al., 2013). Evaluation by human raters can be conducted through system comparisons, rankings (Bojar et al., 2016a), or absolute judgments, direct assessments (Graham et al., 2013). Absolute judgments allow one to efficiently compare a large number of systems. The evaluation of translations as isolated sentences, full paragraphs or documents is also an important factor in the cost/quality trade-offs (Carpuat and Simard, 2012). Isolated sentence evaluation is generally more efficient but fails to penalize contextual mistakes (Tu et al., 2018; Hardmeier et al., 2015).

Automatic evaluation typically collects human reference translations and relies on an automatic metric to compare human references to system outputs. Automatic metrics typically measure the overlap between references and system outputs. A wide variety of metrics has been proposed, and automated metrics is still an active area of research. BLEU (Papineni et al., 2002) is the most common metric. It measures the geometric average of the precision over hypothesis n-grams with an additional penalty to discourage short translations. NIST (Doddington, 2002) is similar but considers up-weighting rare, informative n-grams. TER (Snover et al., 2006) measures an edit distance, as a way to estimate the amount of work to post-edit the hypothesis into the reference. METEOR (Banerjee and Lavie, 2005) suggested rewarding n-gram beyond exact matches, considering synonyms. Others are proposing to use contextualized word embeddings, like BERTscore (Zhang et al., 2019). Rewarding multiple alternative formulations is also the primary motivation behind multiple-reference based evaluation (Nießen et al., 2000). Dreyer and Marcu (2012) introduced an annotation tool and process that can be used to create meaning-equivalent networks that encode an exponential number of translations for a given sentence. Orthogonal to the number of references, the quality of the reference translations is also essential to the reliability of automated evaluation (Zbib et al., 2013). This topic itself raises the question of human translation assessment, which is beyond the scope of this paper (Moorkens et al., 2018).

Meta-evaluation studies the correlation between human assessments and automatic evaluations (Callison-Burch et al., 2006, 2008; Callison-Burch, 2009). Indeed, automatic evaluation is useful only if it rewards hypotheses perceived as fluent and adequate by a human. Interestingly, previous work (Bojar et al., 2016a) has shown that a higher correlation can be achieved when comparing similar systems than when comparing different types of systems, e.g. phrase-based vs neural vs rule-based. In particular, rule-based systems can be penalized as they produce less common translations, even when such translations are fluent and adequate. Similarly, recent benchmark results comparing neu-
ral systems on high resource languages (Bojar et al., 2018; Barrault et al., 2019) have shown mismatches between the systems with highest BLEU score and the systems faring the best in human evaluations. Freitag et al. (2019); Edunov et al. (2019) study this mismatch in the context of systems trained with back-translation (Sennrich et al., 2016) and noisy back-translation (Edunov et al., 2018). They observe that systems training with or without back-translation (BT) can reach a similar level of overlap (BLEU) with the reference, but hypotheses from BT systems are more fluent, both measured by humans and by a language model (LM). They suggest considering LM scores in addition to BLEU.

Freitag et al. (2019); Edunov et al. (2019) point at translationese as a major source of mismatch between BLEU and human evaluation. Translationese refers to artifacts from the source language present in the translations, i.e. human translations are often less fluent than natural target sentences due to word order and lexical choices influenced by the source language (Koppel and Ordan, 2011). The impact of translationese on evaluation has recently received attention (Toral et al., 2018; Zhang and Toral, 2019; Graham et al., 2019). In the present work, we are specifically concerned that the presence of translationese in the references might cause overlap-based metrics to reward hypotheses with translationese language more than hypotheses using more natural language. The question of bias to a specific reference has also been raised in the case of monolingual human evaluation (Fomicheva and Specia, 2016; Ma et al., 2017). The impact of translationese in test sets is related to but different from the impact of translationese in the training data (Kurokawa et al., 2009; Lembersky et al., 2012; Bogoychev and Sennrich, 2019; Riley et al., 2019).

In this work, we explore collecting a single reference translation, using human paraphrases to steer away as much as possible from biases in the reference translation that affect the automatic metrics to prefer MT output with the same style (e.g. translationese). Automatic methods to extract paraphrase n-grams (Zhou et al., 2006) or full sentence paraphrases (Kauchak and Barzilay, 2006; Bawden et al., 2020; Thompson and Post, 2020) have been used to consider multiple references. In contrast, we generate a single unbiased reference translation generated by humans instead of trying to cover a wider space of possible translations. In contrast to human paraphrasing (our instructions asked for most diverse paraphrases), automatic paraphrasing are still far from perfect (Roy and Grangier, 2019) and mostly generate local changes that do not steer away from biases as e.g. introducing different sentence structures.

3 Collecting High Quality and Diverse References

We acquired two types of new reference translations: first, we asked a professional translation service to provide an additional reference translation. Second, we used the same service to paraphrase existing references, asking a different set of linguists.

3.1 Additional Standard References

We asked a professional translation service to create additional high quality references to measure the effect of different reference translations. The work was equally shared by 10 professional linguists. The use of CAT tools (dictionaries, translation memory, MT) was specifically disallowed, and the translation service employed a tool to disable copying from the source field and pasting anything into the target field. The translations were produced by experienced linguists who are native speakers in the target language. The original WMT English→German newstest2019 reference translations have been generated in sequence while keeping an 1-1 alignment between sentences. This should help the linguists to use some kind of document context. We instead shuffled the sentences to also get translations from different linguists within a document and avoid systematic biases within a document. The collection of additional references not only may yield better references, but also allows us to conduct various types of multi-reference evaluation. In addition of applying multi-reference BLEU, it also allows us to select the most adequate option among the alternative references for each sentence, composing a higher quality set.

3.2 Diversified Paraphrased References

The product of human translation is assumed to be ontologically different from natural texts (Koppel and Ordan, 2011) and is therefore often called translationese (Gellerstam, 1986). Translationese includes the effects of interference, the process by which the source language leaves distinct marks in the translation, e.g. word order, sentence structure (monotonic translation) or lexical choices. It also often brings simplification (Laviosa, 1997), as
The Bells of St. Martin’s Fall Silent as Churches in Harlem Struggle.

Die Glocken von St. Martin verstummen, da Kirchen in Harlem Probleme haben.

Die Probleme in Harlems Kirchen lassen die Glocken von St. Martin verstummen.

Die Kirchen in Harlem kämpfen mit Problemen, und so läuten die Glocken von St. Martin nicht mehr.

Table 1: Reference examples of a typical translation and two different paraphrases of this translation. The paraphrases are not only very different from the source sentence (e.g. sentence structure), but also differ a lot when compared to each other.

We explore collecting diverse references using paraphrasing to steer away from translationese, with the ultimate goal of generating a natural-to-natural test set, where neither the source sentences nor the reference sentences contain translationese artifacts. In an initial experiment on a sample of 100 sentences, we asked linguists to paraphrase (translated) sentences. The paraphrased references had only minor changes and consequently only minor impact on the automatic metrics. Therefore, we changed the instructions and asked linguists to paraphrase the sentence as much as possible while also suggesting using synonyms and different sentence structures. The paraphrase instructions are shown in Figure 1. These instructions satisfy not only our goal to generate an unbiased sentence, but also have the side effect that two paraphrases of the same sentence are quite different. All our paraphrase experiments in this paper are done with these instructions. One might be concerned that paraphrasing “as much as possible” might yield excessive reformulation at the expense of adequacy in some cases. To compensate for this in the present paper, we collect adequacy ratings for all produced paraphrases. These ratings allow us to select the most adequate paraphrase from among available alternatives for the same sentence, which results in a composite high paraphrase set with strong adequacy ratings (see Table 2). A paraphrase example is given in Table 1. Even without speaking any German, one can easily see that the paraphrases have a different sentence structure than the source sentence, and both paraphrases are quite different.

4 Experimental Set-up

4.1 Data and Models

We use the official submissions of the WMT 2019 English→German news translation task (Barrault et al., 2019) to measure automatic scores for different kinds of references. We then report correlations with the WMT human ratings from the same evaluation campaign. We chose English→German as this track had the most submissions and the outputs
with the highest adequacy ratings.

### 4.2 Human Evaluation

We use the same direct assessment template as was used in the WMT 2019 evaluation campaign. Human raters are asked to assess a given translation by how adequately it expresses the meaning of the corresponding source sentence on an absolute 0-100 rating scale. We acquire 3 ratings per sentence and take the average as the final sentence score. In contrast to WMT, we do not normalize the scores, and report the average absolute ratings.

### 5 Experiments

We generate three additional references for the WMT 2019 English→German news translation task. In addition to acquiring an additional reference (AR), we also asked linguists to paraphrase the existing WMT reference and the AR reference (see Section 3 for details). We refer to these paraphrases as WMT.p and AR.p.

#### 5.1 Human Evaluation of References

It is often believed that the most accurate translations should also yield the highest correlation with humans ratings when used as reference for an automatic metric. For that reason, we run a human evaluation (Section 4.2) for all reference translations to test this hypothesis (Table 2). While all reference translations yield high scores, the paraphrased references are rated as slightly less accurate. We suspect that this may at least in part be an artifact of the rating methodology. Specifically, translations whose word order matches that of the source (i.e. translationese) are easier to rate than translations that use very different sentence structures and phrasing than the source sentence. We generated our paraphrased reference translation with the instructions to modify the translations as much as possible. Therefore, the non-translationese, perhaps more natural, nature of the paraphrased translations make it more demanding to assign an accurate rating.

As a by-product of these ratings, we consider selecting the best rated references among alternatives for each sentence. Representing this method of combining reference sets with the HQ() function, we generate 3 new reference sets. These are (a) HQ(WMT, AR), abbreviated as HQ(R); (b) HQ(WMT.p, AR.p), abbreviated as HQ(P); and (c) HQ(WMT, AR, AR.p, WMT.p), abbreviated as HQ(all 4). Interestingly, the combined paraphrased reference HQ(P) has a higher human rating than WMT or AR alone.

<table>
<thead>
<tr>
<th></th>
<th>adequacy rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT</td>
<td>85.3</td>
</tr>
<tr>
<td>WMT.p</td>
<td>81.8</td>
</tr>
<tr>
<td>AR</td>
<td>86.7</td>
</tr>
<tr>
<td>AR.p</td>
<td>80.8</td>
</tr>
<tr>
<td>HQ(R) [WMT+AR]</td>
<td>92.8</td>
</tr>
<tr>
<td>HQ(P) [WMT.p+AR.p]</td>
<td>89.1</td>
</tr>
<tr>
<td>HQ(all 4) [all 4]</td>
<td>95.3</td>
</tr>
</tbody>
</table>

Table 2: Human adequacy assessments for different kinds of references, over the full set of 1997 sentences.

#### 5.2 Correlation with Human Judgement

Table 3 provides the system-level rank-correlations (Spearman’s ρ and Kendall’s τ) of BLEU (calculated with sacreBLEU (Post, 2018)) evaluating translations of newstest2019 for different references. On the full set of 22 submissions, all 3 new references (AR, WMT.p, AR.p) show higher correlation with human judgment than the original WMT reference, with the paraphrased references WMT.p coming out on top. Furthermore, each paraphrased reference set shows higher correlation when compared to the reference set that it was paraphrased from.

<table>
<thead>
<tr>
<th></th>
<th>ρ</th>
<th>τ</th>
</tr>
</thead>
<tbody>
<tr>
<td>single ref</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMT</td>
<td>0.88</td>
<td>0.72</td>
</tr>
<tr>
<td>AR</td>
<td>0.89</td>
<td>0.76</td>
</tr>
<tr>
<td>WMT.p</td>
<td>0.91</td>
<td>0.79</td>
</tr>
<tr>
<td>AR.p</td>
<td>0.89</td>
<td>0.77</td>
</tr>
<tr>
<td>single ref</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HQ(R)</td>
<td>0.91</td>
<td>0.78</td>
</tr>
<tr>
<td>HQ(P)</td>
<td>0.91</td>
<td>0.78</td>
</tr>
<tr>
<td>HQ(all 4)</td>
<td>0.91</td>
<td>0.79</td>
</tr>
<tr>
<td>multi ref</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR+WMT</td>
<td>0.90</td>
<td>0.75</td>
</tr>
<tr>
<td>AR.p+WMT.p</td>
<td>0.90</td>
<td>0.79</td>
</tr>
<tr>
<td>all 4</td>
<td>0.90</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 3: Spearman’s ρ and Kendall’s τ for the WMT2019 English→German official submissions with human ratings conducted by the WMT organizers.

Although, the combined reference HQ(R) (Section 5.1) improves correlation when compared to the non-paraphrased reference sets (WMT and AR), not one of the three combined references HQ(R),

---

2We used the scipy implementation in all our experiments: [https://docs.scipy.org/doc/scipy/reference/stats.html](https://docs.scipy.org/doc/scipy/reference/stats.html)

3BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+test.wmt19+tok.intl+version.1.4.2
HQ(P), HQ(all 4) shows higher correlation than the paraphrased reference set WMT.p. This result casts doubt on the belief that if references are rated as more adequate, it necessarily implies that such references will yield more reliable automated scores.

We further find that multi-reference BLEU (calculated with sacreBLEU) does not exhibit better correlation with human judgments either than single-reference BLEU or than the composed reference sets HQ(x). It is generally assumed that multi-reference BLEU yields higher correlation with human judgements due to the increased diversity in the reference translations. However, combining two translated reference sets that likely share the same systematic translationese biases does still prefers translationese translations. Interestingly, multi-reference BLEU with multiple paraphrases also does not show higher correlation than single-reference BLEU. Combining all 4 references with multi reference BLEU shows the same correlation numbers as the combination of AR+WMT. As we will see later, the BLEU scores calculated with paraphrased references are much lower than those calculated with standard references. They have fewer n-gram matches, which are mostly only a subset of the n-gram matches of the standard references. Adding paraphrased references to a mix of standard references therefore has a small effect on the total number of n-gram matches, and as a consequence the scores are not much affected.

As a consequence, we looked deeper into the correlation between a subset of the systems that performed best in human evaluation, where correlation for metrics calculated on the standard reference is known to break down. Kendall’s τ rank correlation as a function of the top k systems can be seen in Figure 2. During the WMT 2019 Metric task (Ma et al., 2019), all official submissions (using the original WMT reference) had low correlation scores with human ratings. The paraphrased references improve especially on high quality system output, and every paraphrased reference set (dotted line) outperforms its corresponding unparaphrased set (same-color solid line).

These improvements in ranking can be seen in Table 4, which reports the actual BLEU scores of the top seven submissions with four different references. Since we asked humans to paraphrase the WMT reference as much as possible (Section 3) to get very different sentences, the paraphrased BLEU scores are much lower than what one expects for a high-quality system. Nevertheless, the system outputs are better ranked and show the highest correlation of any references explored in this paper.

<table>
<thead>
<tr>
<th>System</th>
<th>WMT</th>
<th>HQ(R)</th>
<th>WMT.p</th>
<th>HQ(P)</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB</td>
<td>43.6</td>
<td>42.3</td>
<td>15.1</td>
<td>15.0</td>
<td>0.347</td>
</tr>
<tr>
<td>Micr.sd</td>
<td>44.8</td>
<td>42.1</td>
<td>14.9</td>
<td>14.9</td>
<td>0.311</td>
</tr>
<tr>
<td>Micr.dl</td>
<td>44.8</td>
<td>42.2</td>
<td>14.9</td>
<td>14.9</td>
<td>0.296</td>
</tr>
<tr>
<td>MSRA</td>
<td>46.0</td>
<td>42.1</td>
<td>14.2</td>
<td>14.1</td>
<td>0.214</td>
</tr>
<tr>
<td>UCAM</td>
<td>44.1</td>
<td>40.4</td>
<td>14.2</td>
<td>14.2</td>
<td>0.213</td>
</tr>
<tr>
<td>NEU</td>
<td>44.6</td>
<td>40.8</td>
<td>14.0</td>
<td>14.1</td>
<td>0.208</td>
</tr>
<tr>
<td>MLLP</td>
<td>42.4</td>
<td>38.3</td>
<td>13.3</td>
<td>13.4</td>
<td>0.189</td>
</tr>
</tbody>
</table>

Table 4: BLEU scores of the best submissions of WMT2019 English→German.

### 5.3 Alternative Metrics

Any reference-based metric can be used with our new reference translations. In addition to BLEU, we consider TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005), chrF (Popović, 2015), the f-score variant of BERTScore (Zhang et al., 2019) and Yisi-1 (Lo, 2019) (winning system of WMT 2019 English→German metric task). Table 5 compares these metrics. As we saw in Figure 2, the paraphrased version of each reference set yields higher correlation with human evaluation.
across all evaluated metrics than the corresponding original references, with the only exception of TER for HQ(P). Comparing the two paraphrased references, we see that HQ(P) shows higher correlation for chrF and Yisi when compared to WMT.p. In particular Yisi (which is based on word embeddings) seems to benefit from the higher accuracy of the reference translation.

### Table 5: WMT 2019 English → German: Correlations (Kendall’s \( \tau \)) of alternative metrics: BLEU, 1.0 - TER, chrF, METEOR, BERTScore, and Yisi-1.

<table>
<thead>
<tr>
<th>Metric</th>
<th>WMT</th>
<th>HQ(R)</th>
<th>WMT.p</th>
<th>HQ(P)</th>
<th>HQ(all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.72</td>
<td>0.78</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>1 - TER</td>
<td>0.71</td>
<td>0.74</td>
<td>0.71</td>
<td>0.67</td>
<td>0.74</td>
</tr>
<tr>
<td>chrF</td>
<td>0.74</td>
<td>0.81</td>
<td>0.78</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>MET</td>
<td>0.74</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>BERTS</td>
<td>0.78</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.81</td>
</tr>
<tr>
<td>Yisi-1</td>
<td>0.78</td>
<td>0.84</td>
<td>0.84</td>
<td>0.86</td>
<td>0.84</td>
</tr>
</tbody>
</table>

5.4 WMT18

We acquired a paraphrased as-much-as-possible reference (WMT.p) for newstest2018 English → German with the same instruction as used before (Figure 1). The test set newstest2018 is a joint test set which means that half of the sentences have been originally written in English and translated into German, and vice versa. We paraphrased the reference sentences for the forward translated half only as we want to have a natural English source sentence. Correlation with human rankings of the WMT18 evaluation campaign are summarized in Table 6. The paraphrased reference WMT.p show higher correlations with human judgement for all metrics.

### Table 6: WMT 2018 English → German: Kendall’s \( \tau \).

<table>
<thead>
<tr>
<th>Reference</th>
<th>BLEU</th>
<th>chrF</th>
<th>METEOR</th>
<th>BERTS</th>
<th>Yisi-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT</td>
<td>0.75</td>
<td>0.76</td>
<td>0.75</td>
<td>0.80</td>
<td>0.82</td>
</tr>
<tr>
<td>WMT.p</td>
<td>0.91</td>
<td>0.82</td>
<td>0.84</td>
<td>0.90</td>
<td>0.91</td>
</tr>
</tbody>
</table>

6 Why Paraphrases?

While the top WMT submissions use very similar approaches, there are some techniques in MT that are known to produce more natural (less translationese) output than others. We run experiments with a variety of models that have been shown that their actual quality scores have low correlation with automatic metrics. In particular, we focus on back-translation (Sennrich et al., 2016) and Automatic Post Editing (APE, Freitag et al. (2019)) augmented systems trained on WMT 2014 English → German. All these systems have in common that they generate less translationese output, and thus BLEU with translationese references under-estimate their quality. The experiment in this section follows the setup described in Freitag et al. (2019).

We run adequacy evaluation on WMT newstest 2019 for the 3 systems, as described in Section 4.2. Both the APE and the BT models, which use additional target-side monolingual data, are rated higher by humans than the system relying only on bitext. Table 7 summarizes the BLEU scores for our different reference translations. All references generated with human translations (WMT, HQ(R) and HQ(all 4)) show negative correlation with human ratings for these extreme cases and produce the wrong order. On the other hand, all references that rely purely on paraphrased references do produce the correct ranking of these three systems. This further suggests that reference translations based on human translations bias the metrics to generate higher scores for translationese outputs. By paraphrasing the reference translations, we undo this bias, and the metric can measure the true quality of the underlying systems with greater accuracy.

### Table 7: BLEU scores for WMT newstest 2019 English → German for MT systems trained on bitext, augmented with BT or using APE as text naturalizer. The correct column indicates if the model ranking agrees with human judgments.

<table>
<thead>
<tr>
<th>Reference</th>
<th>bitext</th>
<th>APE</th>
<th>BT</th>
<th>correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>84.5</td>
<td>86.1</td>
<td>87.8</td>
<td>✓</td>
</tr>
<tr>
<td>WMT</td>
<td>39.4</td>
<td>34.6</td>
<td>37.9</td>
<td>✗</td>
</tr>
<tr>
<td>WMT.p</td>
<td>12.5</td>
<td>12.7</td>
<td>12.9</td>
<td>✓</td>
</tr>
<tr>
<td>HQ(R)</td>
<td>35.0</td>
<td>32.1</td>
<td>34.9</td>
<td>✗</td>
</tr>
<tr>
<td>HQ(p)</td>
<td>12.4</td>
<td>12.8</td>
<td>13.0</td>
<td>✓</td>
</tr>
<tr>
<td>HQ(all 4)</td>
<td>27.2</td>
<td>25.8</td>
<td>27.5</td>
<td>✗</td>
</tr>
</tbody>
</table>

This finding, that existing reference translation methodology may systematically bias against modeling techniques known to improve human-judged quality, raises the question of whether previous research has incorrectly discarded approaches that actually improved the quality of MT. Releasing all reference translations gives the community a chance to revisit some of their decisions and measure quality differences for high quality systems.
7 Characterizing Paraphrases

7.1 Alignment

One typical characteristic of translationese is that humans prefer to translate a sentence phrase-by-phrase instead of coming up with a different sentence structure, resulting in ‘monotonic’ translations. To measure the monotonicity of the different reference translations, we compute an alignment with fast-align (Dyer et al., 2013) on the WMT 2014 English-German parallel data and compare the alignments of all four references. Table 8 summarizes the average absolute distance of two alignment points for each reference. The paraphrased translations are less monotonic and use a different sentence structure than a pure human translation.

<table>
<thead>
<tr>
<th></th>
<th>WMT</th>
<th>AR</th>
<th>WMT.p</th>
<th>AR.p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.17</td>
<td>5.27</td>
<td>6.43</td>
<td>6.88</td>
</tr>
</tbody>
</table>

Table 8: Average absolute distance per alignment point, as a proxy for word-by-word (‘monotonic’) translation. Lower scores indicate more monotonic translation.

7.2 Matched n-grams

The actual BLEU scores calculated with the paraphrased references are much lower compared to BLEU scores calculated with standard references (Table 4). Nevertheless, the paraphrased references show higher correlation with human judgment, which motivates us to investigate which n-grams of the MT output are actually matching the paraphrased references during BLEU calculation. The n-grams responsible for the most overlap with standard references are generic, common n-grams. In the winning submission of the WMT 2019 English→German evaluation campaign from Facebook, the 4-grams with the highest number of matches are:

- , sagte er . → 28 times (he said.)
- “, sagte er → 14 times (‘, he said)
- fügte hinzu , dass → 8 times (added that)

These matches are crucial to reach high > 40 BLEU scores, and appear in translation when using the same sentence structure as the source sentence. On the other hand, the n-grams overlapping with the paraphrased references show a different picture. They usually reward n-grams that express the semantic meaning of the sentence. The 4-grams with the highest number of matches with the paraphrased references for the same system are:

- Wheeling , West Virginia → 3 times (Wheeling , West Virginia)
- von Christine Blasey Ford → 3 times (from Christine Blasey Ford)
- Erdbeben der Stärke 7,5 → 3 times (7.5 magnitude earthquake)

8 Conclusions

This work presents a study on the impact of reference quality on the reliability of automated evaluation of machine translation. We consider collecting additional human translations as well as generating more diverse and natural references through paraphrasing. We observe that the paraphrased references result in more reliable automated evaluations, i.e. stronger correlation with human evaluation for the submissions of the WMT 2019 English→German evaluation campaign. These findings are confirmed across a wide range of automated metrics, including BLEU, chrF, METEOR, BERTScore and Yisi. We further demonstrate that the paraphrased references correlate especially well for the top submissions of WMT, and additionally are able to correctly distinguish baselines from systems known to produce more natural output (those augmented with either BT or APE), whose quality tends to be underestimated by references with translationese artifacts.

We explore two different approaches to multi-reference evaluation: (a) standard multi-reference BLEU, and (b) selecting the best-rated references for each sentence. Contrary to conventional wisdom, we find that multi-reference BLEU does not exhibit better correlation with human judgments than single-reference BLEU. Combining two standard reference translations by selecting the best rated reference, on the other hand, did increase correlation for the standard reference translations. Nevertheless, the combined paraphrasing references are of higher quality for all techniques when compared to the standard reference counter part.

We suggest using a single paraphrased reference for more reliable automatic evaluation going forward. Although a combined paraphrased reference shows slightly higher correlation for embedding based metrics, it is over twice as expensive to construct such a reference set. To drive this point home, our experiments suggest that standard reference translations may systematically bias against modelling techniques known to improve human-judged quality, raising the question of whether previous
research has incorrectly discarded approaches that actually improved the quality of MT. Releasing all reference translations gives the community a chance to revisit some of their decisions and measure quality differences for high quality systems and modelling techniques that produce more natural or fluent output.

As a closing note, we would like to emphasize that it is more difficult for a human rater to rate a paraphrased translation than a translationese sentence, because the latter may share a similar structure and lexical choice to the source. We suspect that human evaluation is also less reliable for complex translations. Future work can investigate whether finer ratings could correct the bias in favor of lower effort ratings, and how this may interact with document-level evaluation.

References


Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2019. Translationese as a language in "multilingual" nmt.


