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. In particular, language pairs with abundant training data have benefited tremendously from neural machine translation techniques (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, hyper-parameter 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 ranges 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 concentrate in a limited part of the space of target sentences with the same meaning. We show that this part of the space is different from original native target sentences. Human translators tend to generate translation which exhibit *translationese* language, i.e. sentences with source artifacts (Koppel and Ordan, 2011). As a consequence, automatic metrics are biased to pro-

duce higher scores for translationese MT outputs than for more natural outputs. Without additional instructions, we show that collecting different human translations does not produce a rich set of valid translations. This is problematic because collecting only a single style of references fails to reward systems that might produce alternative but equally accurate translations. Because of this lack of diversity, multi-reference evaluations like multi-reference BLEU are also biased to prefer that specific style of translation. We however find that selecting the most adequate translation within a set of alternative references can improve the quality of automated evaluation, albeit not in all cases.

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 paraphrases of reference translations. We show that (human) paraphrases have multiple interesting properties in terms of diversity, accuracy, naturalness and correlation 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[1] a rich set of diverse references to encourage research in systems producing other types of translations, and reward a wider range of generated language.

[1] https://github.com/google/wmt19-paraphrased-references

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. With similar cost motivations, previous work has advocated for contracting evaluation to crowd workers instead of language experts (Goto et al., 2014; Graham et al., 2017). The evaluation of translations as isolated sentences, full paragraphs or documents is also an important factor in the cost/quality tradeoffs (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). 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 neural 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).

Unlike Edunov et al. (2019), we do not circumvent translationese preference biases with language model scores, as we would rather the evaluation to be independent from language modeling choices like the LM architecture or its training distribution. We instead explore collecting more diverse hypotheses, using paraphrases to steer away from translationese. Paraphrases have already been considered for the purpose of MT evaluation. Automatic methods to extract paraphrase n-grams (Zhou et al., 2006) or generate full sentence paraphrases (Kauchak and Barzilay, 2006) have been used to consider multiple references. These strategies however require factoring in the quality of the paraphrasing system in the evaluation, as such systems are still far from perfect (Roy and Grangier, 2019). Previous work has also considered automatic paraphrases for system tuning (Madnani et al., 2007; Marton et al., 2009).

3 Collecting High Quality and Diverse References

In this section, we describe how we acquired additional references. We tried two approaches: 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 Increasing reference quality

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 linguists who are native speakers in the target language and have many years of experience in translation tasks. The translation setup is similar to the reference generation in WMT. On a high level, we could not find any significant differences in the way WMT generated their references for the WMT English→German translation task. Of course, we used a different vendor and the vendors themselves use different quality assessments and different linguists. The collection of additional references not only may yield better references, but also allows us to conduct various types of multi-reference eval-

uation. In addition to traditional approaches like 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, natural references through paraphrasing

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 or lexical choices. It also often brings simplification (Laviosa, 1997), as the translator might impoverish the message, the language, or both. Most importantly for machine translation evaluation, two translations of the same source are very similar and only cover a small part of all possible translations. The troubling implication is that a reference set of translationese sentences is biased to assign higher word overlap scores to MT outputs that produces a similar translationese style, and penalizes MT output with more natural targets (Freitag et al., 2019). Collecting different types of adequate references could therefore uncover alternative high quality systems producing different types of outputs.

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. Paraphrased references therefore cover a wider diversity of target sentences than the traditional translations, which we quantify in Section 7.3. All our paraphrase experiments in this paper are done with these instructions. As a side note, one might be concerned that paraphrasing "as much as possible" might yield excessive reformulation at the expense of adequacy in some cases. It may indeed be true that more investigation into the manner of paraphrasing would yield better instructions. 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 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 that both paraphrases are quite different from each other.

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.

Task: Paraphrase the sentence as much as possible:

To paraphrase a source, you have to rewrite a sentence without changing the meaning of the original sentence.

1. Read the sentence several times to fully understand the meaning
2. Note down key concepts
3. Write your version of the text without looking at the original
4. Compare your paraphrased text with the original and make minor adjustments to phrases that remain too similar

Please try to change as much as you can without changing the meaning of the original sentence. Some suggestions:

1. Start your first sentence at a different point from that of the original source (if possible)
2. Use as many synonyms as possible
3. Change the sentence structure (if possible)

Figure 1: Instructions used to paraphrase an existing translation *as much as possible*.

Source	The Bells of St. Martin's Fall Silent as Churches in Harlem Struggle .
Translation	Die Glocken von St. Martin verstummen , da Kirchen in Harlem Probleme haben .
Paraphrase	Die Probleme in Harlems Kirchen lassen die Glocken von St. Martin verstummen .
Paraphrase	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.

5.1 Human Evaluation of References

For the purpose of MT evaluation, a reference translation is considered good if its use as reference for an automated metric yields scores that can replace human judgments. In other words, measuring overlap between an MT output and a good reference correlates well with human ratings of that MT output. It is often believed that reference translations with high human ratings should also be good for automated evaluation. For that reason, we run a quality human evaluation (see Section 4.2) for all our four reference translations to test this hypothesis. Table 2 summarizes the average human scores for all references.

While all four 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 original source sentence, because it is easier for a rater to compare them word-by-word. We generated our paraphrased reference translation with the instructions to modify the translations as much as possible, using different wordings and a different sentence structure. Therefore, the non-translationese, perhaps more natural, nature of the paraphrased translations make it more demanding to assign an accurate adequacy rating. In future work, we want to investigate if finer ratings could correct the bias in favor of lower effort ratings.

	adequacy rating
WMT	85.3
WMT.p	81.8
AR	86.7
AR.p	80.8
HQ(R) [WMT+AR]	92.8
HQ(P) [WMT.p+AR.p]	89.1
HQ(all 4) [all 4]	95.3

Table 2: Human adequacy assessments for different kinds of references, over the full set of 1997 sentences. HQ(P) has been generated by picking sentence-by-sentence the more accurate rated translation from WMT.p and AR.p. HQ(R) and HQ(all 4) have been generated with the same method by either combining WMT and AR or all four reference translations.

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.

5.2 Correlation with Human Judgement

Table 3 provides the rank-correlations (Spearman's ρ and Kendall's τ)[2] of BLEU[3] 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 "standard" reference set that it was paraphrased from.

By combining two reference translations by using the reference translation with the higher human rating (See 5.1), we generated reference translations which are rated as more accurate. Although this approach improves correlation when applied to the non-paraphrased reference sets (WMT and AR), not one of the three combined references HQ(R), 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.

The other standard approach to using multiple references is multi-reference BLEU. We find that multi-reference BLEU 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 (i.e. WMT and AR) does not yield a very diverse set (see Section 7.3). More importantly, measuring overlap with an extra translationese reference will not reward natural language more. 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 the 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 significantly affected.

Full Set (22)	Reference	ρ	τ
single ref	WMT	0.88	0.72
	AR	0.89	0.76
	WMT.p	**0.91**	**0.79**
	AR.p	0.89	0.77
single ref	HQ(R)	**0.91**	0.78
	HQ(P)	**0.91**	0.78
	HQ(all 4)	**0.91**	**0.79**
multi ref	AR+WMT	0.90	0.75
	AR.p+WMT.p	0.90	**0.79**
	all 4	0.90	0.75

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

Note that the correlation numbers already appear relatively high for the full set of systems. This is because both Kendall's τ and Spearman's ρ rank correlation operate over all possible pairs of systems. Since the submissions to WMT2019 covered a wide range of translation qualities, any metric able to distinguish the highest-scoring and lowest-scoring systems will already have a high correlation. Therefore, small numeric increases as demonstrated in Table 3 can correspond to much larger improvements in the local ranking of systems.

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 ref-

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

[3]BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+test.wmt19+tok.intl+version.1.4.2

erence 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). Interestingly, WMT.p shows higher correlation than HQ(P) when looking only at top submissions. Both our paraphrased reference WMT.p and AR.p, produced the correct order for the top seven submissions.

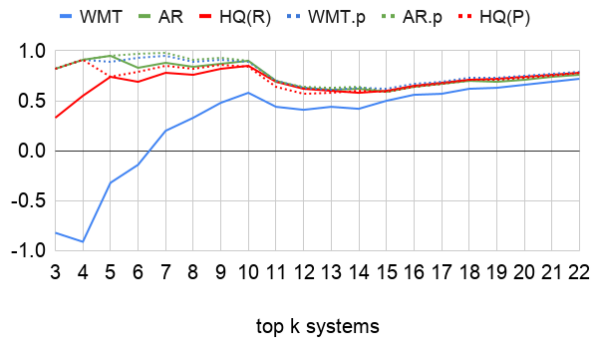

Figure 2: Kendall's τ correlation of BLEU for the best k systems (based on human ratings).

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.

	WMT	HQ(R)	WMT.p	HQ(P)	human
FB	43.6	**42.3**	**15.1**	**15.0**	**0.347**
Micr.sd	44.8	42.1	14.9	14.9	0.311
Micr.dl	44.8	42.2	14.9	14.9	0.296
MSRA	**46.0**	42.1	14.2	14.1	0.214
UCAM	44.1	40.4	14.2	14.2	0.213
NEU	44.6	40.8	14.0	14.1	0.208
MLLP	42.4	38.3	13.3	13.4	0.189

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.

metric	WMT	HQ(R)	WMT.p	HQ(P)	HQ(all)
BLEU	0.72	0.78	**0.79**	**0.79**	**0.79**
1 - TER	0.71	**0.74**	0.71	0.67	**0.74**
chrF	0.74	0.81	0.78	**0.82**	0.78
MET	0.74	**0.81**	**0.81**	**0.81**	0.80
BERTS	0.78	**0.82**	**0.82**	**0.82**	0.81
Yisi-1	0.78	0.84	0.84	**0.86**	0.84

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

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) for data and models.

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 6 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.

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

Table 6: 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.

This finding, that existing reference translation methodology may systematically bias against modelling 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 vs. Translations

We examine in more detail the characteristics of original text, translations and 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 7 summarizes the average absolute distance of two alignment points for each reference translation. As expected, the paraphrased translations are less monotonic and use a different sentence structure than a pure human translation.

Reference	Average distance
WMT	5.17
AR	5.27
WMT.p	6.43
AR.p	6.88

Table 7: Average absolute distance per alignment point, as a proxy for measuring 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 (see 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 German n-grams. In the winning submission of the WMT 2019 English→German evaluation campaign from Facebook, the 4grams that have 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*)

One should note that this effect is related to the down-weighting of common language in favor of more informative content, which has been experimented in the past with explicit term weighting (Babych and Hartley, 2004; Wong and Kit, 2011).

7.3 Round-trip translation study

We assess the following hypotheses: (i) translations of the same sentence tend to be similar to each other, i.e. they concentrate in a small part of the target sentence space, and (ii) a target sentence and a paraphrase tend to be further from each other, i.e. paraphrases allow access to a wider variety of target sentences with the same meaning. In this study, we also concentrate on English as the source language and German as the target language.

For this experiment, we need an English source sentence along with a corresponding *original* German sentence. Unfortunately, it is impossible to have a pair of corresponding English and German sentences in which both sides are original (i.e. non translated). We therefore devise a compromise with an artificial source obtained through translation, which we call "en.tr". We are aware of the drawback of translated sources (Bogoychev and Sennrich, 2019) but this is unfortunately the only way to have a German original target sentence to refer to. This experiment relies on 100 German news sentences randomly sampled from German→English newstest2019.

We task a professional translation service to create the English source (en.tr) from the German original sentence (de.orig). From the English source (en.tr), we rely on the same service to create two translations (de.tr1 and de.tr2). We rely on the same service again to create two paraphrases from the first translation (de.tr1.p1 and de.tr1.p2), and two paraphrases of the original German sentences (de.orig.p1 and de.orig.p2). This process is illustrated in Figure 3. Each linguist was only allowed to work on one of these 7 reference generations and each task has been processed by 2 humans (50 sentences each).

At each step in the process, we task annotators to validate the adequacy of the translations and paraphrases. Table 8 shows high adequacy for all translations. It also shows that paraphrases tend to be judged less adequate than the original and translation. We also observe higher variance and less inter-annotator agreement for paraphrases. This likely indicates that their ratings involve more work than the rating of translationese with simpler source correspondence (translationese tends to have similar sentence structure as the source sentence). Overall, it seems that some raters have difficulty assigning good scores for correct translations with different sentence structure than the source sentence. We want to confirm this in future work and come up with a human evaluation setup that is unbiased by the sentence structure.

	adequacy rating
de.orig	90.6
de.tr1	90.4
de.tr2	89.9
de.orig.p1	80.5
de.orig.p2	76.9
de.tr1.p1	78.3
de.tr1.p2	85.0

Table 8: Human adequacy assessments for different kinds of references, over the random sample of 100 sentences.

Table 9 reports BLEU by comparing all pairs of the 100-sentence sets created through this process, as a proxy for understanding how similar these domains are. These results verify our hypotheses: translations (de.tr1 and de.tr2) are the most similar pairs ($>$ 43 BLEU), while their similarity with the original sentence is much less (27.5 and 24.8 BLEU resp.). This highlights that direct translations tend to concentrate into similar parts of the translation space. For automatic MT evaluation, this implies that systems are currently required to produce translations from a limited part of the space to achieve high BLEU. For instance, if one imagines that de.tr1 is a reference and that de.tr2 and de.orig are systems, BLEU scores will determine that de.tr2 is a far better translation (43.9 BLEU) than the original German sentence (27.4 BLEU). Unsurprisingly, this disagrees with our human adequacy ratings (Table. 8).

The space of valid equivalent target sentences is however much richer than the space of direct

Figure 3: Labeling process to assess the diversity of references for an English-German translation task. P means that linguists paraphrased each sentence *as much as possible* into the same language. T means that linguists translated the sentence from English/German into the other language.

	de.orig	de.tr1	de.tr2	de.orig.p1	de.orig.p2	de.tr1.p1	de.tr1.p2
de.orig		27.5	24.8	21.0	17.0	8.4	15.4
de.tr1	27.4		43.9	16.1	12.4	14.5	22.2
de.tr2	24.7	44.0		15.8	12.7	10.5	20.2
de.orig.p1	20.9	16.1	15.8		22.4	7.5	21.0
de.orig.p2	16.9	12.3	12.7	22.3		10.2	17.2
de.tr1.p1	8.4	14.4	10.5	7.5	10.2		11.7
de.tr1.p2	15.4	22.2	20.2	21.0	17.2	11.7	

Table 9: BLEU scores comparing different reference generation approaches. All numbers are calculated using one reference as hypothesis (column) and another reference as reference (row) in sacrebleu. The corresponding BLEU scores indicate how similar two references are (higher means more similar).

translations, as shown by the overlap between paraphrases with the original German sentence. The BLEU scores between paraphrases and the original sentence (de.orig) range from 8.4 to 21.0, indicating that this is a rich, diverse set of sentences. In concrete terms, imagine that de.tr2 is the reference translation, and we are comparing systems producing either de.tr1 (typical translationese) or de.tr1.p1 (the same output but made more natural). Although the translationese system does have a somewhat higher accuracy (89.9% vs 78.3%), the BLEU difference exaggerates this difference to a comical extent (43.9 vs. 10.5).

Our experiments also show that paraphrasing is not a silver bullet against translationese effects. Paraphrases tend to be more similar to the sentence they originate from (de.tr1 or de.orig) than to the other German sentences (de.orig, de.tr1 or de.tr2). In other words, there is also a form of language bias leaking from the paraphrased sentence into the paraphrase, which is not unlike source language artifacts appearing in target language (translationese). Unlike translations, however, the set of sentences produced by paraphrasing is not clustered within a very small part of the target space.

7.4 Measuring Translationese

Translationese tends to be simpler, more standardised and more explicit (Baker et al., 1993) compared to original text and can retain typical characteristics of the source language (Toury, 2012). Toral (2019) proposed metrics attempting to quantify the degree of translationese present in a translation. Following their work, we quantify lexical simplicity with two metrics: lexical variety and lexical density. We also calculate the length variety to measure interference from the source.

7.4.1 Lexical Variety

An output is simpler and therefore more translationese when it uses a lower number of unique tokens/words.

$$lex\_variety = \frac{number\ of\ types}{number\ of\ tokens} \quad (1)$$

7.4.2 Lexical Density

Scarpa (2006) found that translationese tends to be lexically simpler and have a lower percentage of content words (adverbs, adjectives, nouns and verbs) compared to original written text.

$$lex\_density = \frac{number\ of\ content\ words}{number\ of\ total\ words} \quad (2)$$

7.4.3 Length Variety

Both MT and humans tend to avoid restructuring the source sentence and stick to sentence structures popular in the source language. This results in a translation with similar length to that of the source sentence. By measuring the length variety, we measure interference in the translation because its length is guided by the source sentence's structure. We compute the normalized absolute length

difference at the sentence level and average the scores over the test set of source-target pairs (x, y):

$$len\_variety = \frac{||x| - |y||}{|x|} \quad (3)$$

Numbers for all three translationese metrics can be found in Table 10. For all metrics, de.tr gets the lowest scores, confirming that standard human translations yield more translationese style output. The paraphrases, on the other hand, have lexical density and length variety that is much higher than both the translated sentences and the original German sentences, though they have a lower lexical variety. This demonstrates that we were able to remove many of the translationese artifacts by paraphrasing as much as possible.

	Lex. Var.	Lex. Density	Len. Var.
de.orig	**0.534**	**0.398**	**0.134**
de.tr	0.509 *-4.6%*	0.391 *-1.8%*	0.131 *-2.2%*
de.orig.p	0.513 *-3.9%*	0.408 ***+2.0%***	0.195 ***+45%***
de.tr1.p	0.522 *-2.2%*	0.400 ***+0.5%***	0.196 ***+46%***

Table 10: Measuring the degree of translationese, reporting percent difference wrt. to de.orig. Higher lexical variety, lexical density, and length variety imply less translationese sentences. Values at or exceeding those of natural text are bolded.

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

Bogdan Babych and Anthony Hartley. 2004. Extending the BLEU MT Evaluation Method with Frequency Weightings. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, page 621. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. *Corpus Linguistics and Translation Studies: Implications and Applications*, chapter 2. John Benjamins Publishing Company, Netherlands.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation. *arXiv preprint arXiv:1911.03362*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016a. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016b. Ten Years of WMT Evaluation Campaigns: Lessons Learnt. *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, page 27.

Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 70–106, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Marine Carpuat and Michel Simard. 2012. The Trouble with SMT Consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2019. On The Evaluation of Machine Translation Systems Trained With Back-Translation. *arXiv preprint arXiv:1908.05204*.

Marina Fomicheva and Lucia Specia. 2016. Reference Bias in Monolingual Machine Translation Evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 77–82, Berlin, Germany. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at Scale and Its Implications on MT Evaluation Biases. In *Proceedings of the Fourth Conference on*

Machine Translation, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, and Yann N. Dauphin. 2017. A Convolutional Encoder Model for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 123–135.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, page 8895. CWK Gleerup.

Shinsuke Goto, Donghui Lin, and Toru Ishida. 2014. Crowdsourcing for Evaluating Machine Translation Quality. In *LREC*, volume 2014, pages 3456–346.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in Machine Translation Evaluation. *CoRR*, abs/1906.09833.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, USA. Association for Computational Linguistics.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1318–1326, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, pages 81–88.

Sara Laviosa. 1997. How comparable can'comparable corpora'be? *Target. International Journal of Translation Studies*, 9(2):289–319.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 255–265, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chi-kiu Lo. 2019. Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513.

Qingsong Ma, Yvette Graham, Timothy Baldwin, and Qun Liu. 2017. Further investigation into reference bias in monolingual evaluation of machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2476–2485, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie J Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127. Association for Computational Linguistics.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 381–390, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. 2018. *Translation Quality Assessment: From Principles to Practice*. Springer.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

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*, pages 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*, pages 392–395.

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

Aurko Roy and David Grangier. 2019. Unsupervised paraphrasing without translation. In *ACL (1)*, pages 6033–6039. Association for Computational Linguistics.

Federica Scarpa. 2006. Corpus-based quality-assessment of specialist translation: A study using parallel and comparable corpora in english and italian. *Insights into specialized translation–linguistics insights. Bern: Peter Lang*, pages 155–172.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Antonio Toral. 2019. Post-editese: an exacerbated translationese. *CoRR*, abs/1907.00900.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Belgium, Brussels. Association for Computational Linguistics.

Gideon Toury. 2012. *Descriptive translation studies and beyond: Revised edition*, volume 100. John Benjamins Publishing.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

John S. White. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Further Approaches. In *Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas*, pages 193–205.

Billy Wong and Chunyu Kit. 2011. Comparative evaluation of term informativeness measures for machine translation evaluation metrics. In *MT Summit*, volume 2011, pages 537–544.

Rabih Zbib, Gretchen Markiewicz, Spyros Matsoukas, Richard Schwartz, and John Makhoul. 2013. Systematic comparison of professional and crowdsourced reference translations for machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 612–616, Atlanta, Georgia. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. *CoRR*, abs/1906.08069.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *Arxiv*, 1904.09675.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 77–84, Sydney, Australia. Association for Computational Linguistics.