Statistical Power and Translationese in Machine Translation Evaluation

Yvette Graham
ADAPT Centre
Trinity College Dublin
ygraham@tcd.ie

Barry Haddow
School of Informatics
University of Edinburgh
bhaddow@inf.ed.ac.uk

Philipp Koehn
Dept of Computer Science
Johns Hopkins University
phi@jhu.edu

Abstract

The term *translationese* has been used to describe features of translated text, and in this paper, we provide detailed analysis of potential adverse effects of translationese on machine translation evaluation. Our analysis shows differences in conclusions drawn from evaluations that include translationese in test data compared to experiments that tested only with text originally composed in that language. For this reason we recommend that reverse-created test data be omitted from future machine translation test sets. In addition, we provide a re-evaluation of a past machine translation evaluation claiming human-parity of MT. One important issue not previously considered is statistical power of significance tests applied to comparison of human and machine translation. Since the very aim of past evaluations was the investigation of ties between human and MT systems, power analysis is of particular importance, to avoid, for example, claims of human parity simply corresponding to Type II error resulting from the application of a low powered test. We provide detailed analysis of tests used in such evaluations to provide an indication of a suitable minimum sample size for future studies.

1 Introduction

Human-translated text is thought to display features that deviate to some degree from those of text originally composed in that language. Baker (1993) report that translated text can: be more explicit than the original source, less ambiguous, simplified (lexical, syntactically and stylistically); display a preference for conventional grammaticality; avoid repetition; exaggerate target language features; as well as display features of the source language. The term *translationese* is often used to describe the presence of such phenomena in translated text.

Standard evaluation protocol in Machine Translation (MT) comprises system tests on a sample of human-translated text. Since creating this human-translated text is expensive, re-use of test sets for both directions of translation is commonplace, without regard to whether source or target contain features of translationese. For example, translation shared tasks at the Conference on Machine Translation (WMT) (Bojar et al., 2018) generally test translation between a given language pair with two portions of data combined to make up the overall test set. Portion (a) of the test data (accounting for approximately 50% of sentences) is made up of text that originated in Chinese that was human-translated into English, while portion (b) (i.e. the remaining 50%), was translated in the opposite direction, originating in English with manual translation into Chinese. The motivation for creating the test data in this way is to create test sets for both directions simultaneously (so at no extra cost).[1]

Although translationese has been cited as a likely confound in MT evaluation results in the past (Lambersky et al., 2012; Toral et al., 2018; Läubli et al., 2018), to the best of our knowledge, no detailed investigation into the impact of translationese on the accuracy of MT evaluation has been reported to date. With this aim, we examine the degree to which translationese phenomena may impact human and automatic evaluation results in MT. We firstly examine past results of WMT shared tasks, a main venue for MT evaluation, and reveal that although system rankings are overall very similar for human evaluation of forward and reverse test data, in a small number of cases system rankings diverge to a more serious degree. For example, for Turkish-English translation at WMT-18 forward and reverse system rankings correlate at only $r = 0.703$ in one case. Apart from human evaluation,

[1] WMT news task ceased employing reverse-created test data in 2019, motivated by the analysis provided in this current work published in an earlier archival version (Graham et al., 2019).

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 72–81,
November 16–20, 2020.

much more concerning is the divergence in forward and reverse rankings when BLEU is relied upon for evaluation of systems, where the correlation can be as low as 0.106 in the worst case.

Subsequently, we provide a reassessment of a human evaluation previously criticized for including reverse-created test data that claimed human parity of Chinese to English MT. We reveal insights into additional potential sources of inaccuracy of conclusions beyond the presence of translationese with the aim of preventing future inaccuracies.

2 Related Work

Hassan et al. (2018) provide one of the earliest claims in MT of systems achieving human-parity in terms of the quality of translations. The reliability of these claims was quickly contested in follow-up studies by Läubli et al. (2018) and Toral et al. (2018), who both drew attention to the 50/50 set-up of test data creation, highlighting the inclusion of reverse-created test data as a likely confound. In their repeat of the human evaluation of the translations produced by Hassan et al. (2018), both Läubli et al. (2018) and Toral et al. (2018) used only test data that originated in the source language.

Inspired by this work, other authors considered the effect of the 50/50 set-up on evaluation using WMT data. Edunov et al. (2019) questioned whether improvements in performance due to back-translation were just an artifact of the test set construction. They found that, whilst back-translation had a disproportionately large positive effect on BLEU for reverse-created test sets, human evaluation showed that back-translation did indeed provide robust improvements to MT for forward-created text. Related to this, Freitag et al. (2019) also showed BLEU to be misleading on the reverse-created part of the test sets, when analysing why their automatic post-editing (APE) method produced improved translations according to human evaluation, but not according to BLEU. Given the concern in the community about using reverse-created test sets, the organisers of the WMT19 news translation task used only forward-created sentences in all their test sets (Barrault et al., 2019). In this current paper we provide detailed evidence to justify this decision.

We note that Zhang and Toral (2019) also provide analysis of the effect of reverse-created test sets on WMT evaluation campaigns. However they focus only on the effect of translationese with respect to human evaluation, without considering its differing effect on automatic evaluation. Also, they do not consider the problem of statistical power in human evaluation, which we raise below.

The use of reverse-created test sets was not the only concern raised by Läubli et al. (2018) and Toral et al. (2018). Both used more context than the original sentence-level evaluation in Hassan et al. (2018), Läubli et al. (2018) now asking human judges to assess entire documents, and Toral et al. (2018) involving assessment of MT output sentences in the order that they appeared in original documents. Furthermore, in contrast to the use of Direct Assessment (Graham et al., 2016) by Hassan et al. (2018), both reassessments used relative ranking, a method formerly used in WMT for evaluation (Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015, 2016), but now abandoned, partly due to low inter-annotator agreement.

Therefore, although both re-evaluations improved the methodology employed in two respects, by eliminating reverse-created test data and including more context, both potentially include other sources of inaccuracy, such as lack of reliability of human judges when human evaluation takes the form of relative ranking.

Furthermore, Toral et al. (2018) employ Trueskill to reach the conclusion that the MT system in question has not achieved human performance, and although Trueskill has been used in past WMT evaluations to produce system rankings, its aim is to minimize the number of judgments required to produce those rankings when resources are limited. So results may not be directly comparable with results of standard statistical significance tests, now current practice at WMT evaluations.

Finally, neither Toral et al. (2018) nor Läubli et al. (2018) discuss statistical power of significance tests used to distinguish the performance of system and human, an important aspect of evaluation and one of particular importance with respect to evaluations that aim to investigate claims of human parity, where Type II error could result in false claims.

Besides criticisms already made of the human evaluation in Hassan et al. (2018), an additional aspect of importance not yet highlighted is the proportion of distinct translations that were included in the original human-parity evaluation of systems, a consideration that also relates strongly to the ques-

tion of statistical power. In most MT human evaluations, it is not feasible to evaluate the full test set of sentences for all systems and it is common to instead evaluate a *sample of translations*, usually drawn at random from the test data. In current WMT evaluations, for example, translations of all test sentences produced by all participating systems are pooled and a random sample is human-evaluated. This method ensures that as great a number as possible of *distinct test sentences* are examined. Alongside system performance estimates, WMT also reports the number of distinct test sentences evaluated, n, and it is this number that they consider the *sample size* used for statistical significance tests subsequently used to draw conclusions about which competing systems outperform others. For example, all else being equal, a difference in system performance estimates for a pair of systems computed from a *larger set of distinct translations* is interpreted as *more reliable*.

	Ave.	z	n	N	System
FWD	67.1	0.185	92	828	Reference-HT
	64.8	0.048	92	828	Combo-5
	64.3	0.042	92	828	Combo-6
	64.3	0.023	92	828	Combo-4
	64.1	0.020	92	828	Reference-PE
	61.1	−0.144	92	828	Reference-WMT
	56.2	−0.345	92	828	Sogou
	50.9	−0.580	92	828	Online-A-1710
	48.5	−0.717	92	828	Online-B-1710
REV	73.8	0.434	89	801	Combo-6
	73.2	0.393	89	802	Combo-5
	72.8	0.392	89	801	Combo-4
	70.3	0.256	89	801	Reference-PE
	70.0	0.252	89	801	Reference-HT
	68.8	0.167	89	801	Sogou
	63.0	−0.089	89	801	Reference-WMT
	60.0	−0.214	89	801	Online-B-1710
	61.1	−0.217	89	802	Online-A-1710
BOTH	69.0	0.235	181	1,629	Combo-6
	68.5	0.218	181	1,629	Reference-HT
	68.9	0.218	181	1,630	Combo-5
	68.5	0.204	181	1,629	Combo-4
	67.1	0.136	181	1,629	Reference-PE
	62.4	−0.093	181	1,629	Sogou
	62.0	−0.117	181	1,629	Reference-WMT
	55.9	−0.402	181	1,630	Online-A-1710
	54.1	−0.469	181	1,629	Online-B-1710

Table 1: Results of Hassan et al. (2018) for forward, reverse and both test set creation directions. N = number of human judgments; n = number of distinct translations, Reference-HT = human translations created by (Hassan et al., 2018), Reference-PE = post-edited online MT system; Reference-WMT = original WMT reference translations; horizontal lines denote clusters according to Wilcoxon rank sum test at $p < 0.05$.

Other MT human evaluations, despite claims of following WMT human evaluation methodology, have diverged from this method of sample size computation, however, including the human-parity evaluation of Hassan et al. (2018) and Läubli et al. (2018). For example, although a large sample of human judgments is reported as $n \geq$ 1,827 per system in Hassan et al. (2018), firstly this number in fact included quality control check translations, generally removed from data before computing sample sizes. More importantly however, very high numbers of repeat evaluations of the same translations were included in the human-parity evaluation of Hassan et al. (2018). In other words, a very low number of distinct test sentences were in fact human evaluated despite reporting a large sample size. The method of computing sample size therefore diverges from that reported of WMT evaluations in a small but important way. The sample size reported instead corresponds to the total number of human ratings collected as opposed to distinct test sentences (as in WMT evaluations). In this current work, we make this important distinction explicit by referring to the number of distinct test sentences evaluated as n and the number of human judgments collected as N. We also recommend this distinction be made and adopted as common practice in future human evaluations of MT or that the number of distinct translations (as opposed to the number of human evaluations) be reported as the sample size.

Table 1 shows results reproduced from the Hassan et al. (2018) data set, where we now report both the number of human judgments collected, N, and the number of distinct test sentences included, n, in addition to adding separate results for forward and reverse-created test data. Only when tested on the less legitimate reverse direction data does MT now appear to outperform human translation. Nonetheless, when interpreting results in Table 1, it is important to remember, however, that the reliability of even the conclusions drawn from forward-created test data only is still uncertain however, due to the small n, as only 92 distinct translations were in fact included in the evaluation claiming human parity. It remains a possibility that, for example, had the number of distinct test sentences evaluated been higher, distinct conclusions would also be drawn.

We therefore rerun the evaluation using the original translation data included in Hassan et al. (2018) with entirely up-to-date WMT human evaluation methodology in addition to ensuring that a suf-

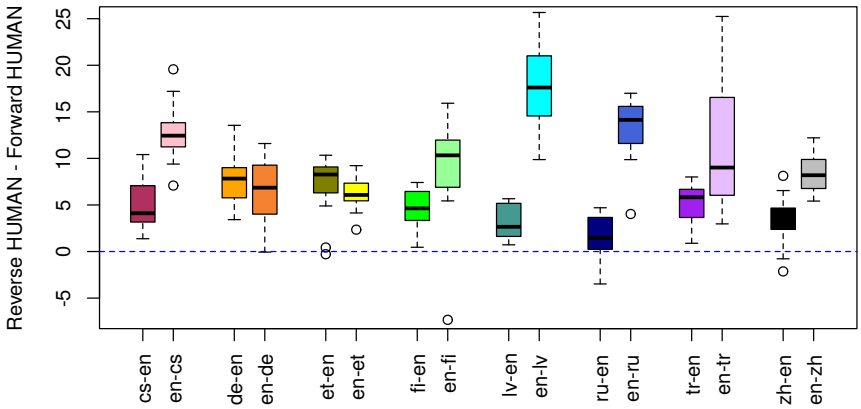

Figure 1: Differences in human evaluation Direct Assessment (DA) scores for test sentences created in the reverse direction to testing and those created in the same/forward direction to testing broken down by language pair, showing that reverse human evaluation scores higher than forward ones in almost all cases.

ficiently large sample of distinct translations are assessed by human judges. We also take into account the very legitimate criticism made by both Toral et al. (2018) and Läubli et al. (2018) and include document-level context in the human evaluation. Furthermore, since no previous evaluation has included statistical power analysis, prior to running our own human evaluation, we examine the power of significance tests to estimate a suitable sample size to decrease the likelihood of Type II error leading to conclusions of human parity due to the application of a low powered test.

Additionally, we examine potential issues for MT evaluation when test data created in the reverse direction to testing is included. Despite being identified by Toral et al. (2018) and Läubli et al. (2018) as a serious cause of concern in MT evaluations, to the best of our knowledge no previous study exists that examines in detail the degree to which reverse-created test data may have skewed past results. The sections that follow therefore include an investigation into the issue of translationese in MT evaluation, in addition to a re-evaluation of Hassan et al. (2018) data with all potential sources of criticism, in terms of test data and evaluation methodology, now taken into account and corrected.

3 Translationese

Using reverse-created test data is thought to unrealistically decrease the difficulty of MT evaluations (Toral et al., 2018; Läubli et al., 2018), because in real-world MT scenarios, input text is unlikely to very often comprise text that has already been translated from the target language. We therefore compare results of systems when test data is split according to the creation direction and examine differences in scores for systems in terms of both human and automatic metrics.

3.1 Human Evaluation

In order to examine differences in human evaluation results when translationese is in test data, we firstly examine WMT-17 and WMT-18 systems and compute two separate human evaluation scores for each system. For each individual system, we compute its *forward Direct Assessment (DA) score*, comprising the average DA score computed only for test sentences that were created in the *same direction as testing*, and a corresponding *reverse DA score* from test data created in the *opposite direction to testing*. Then, to examine the extremity to which MT human evaluation results may differ when systems are tested in the reverse as opposed to forward direction, we subtract a given system's forward DA score (expected to be lower than its reverse counterpart) from its reverse DA score (expected to be higher than its forward counterpart). This provides the difference in human DA scores for each system, with positive differences expected in general since reverse-created test data is hypothesised to be an artificially easier test for MT systems.

Figure 1 shows the distribution of DA score differences (reverse DA − forward DA) for all sys-

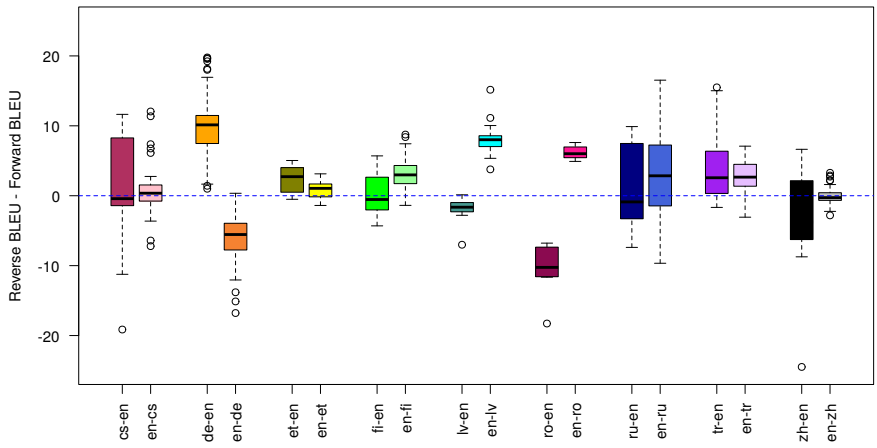

Figure 2: Differences in BLEU scores for systems participating in WMT-15–WMT-18 news translation task computed for test sentences created in the reverse direction to testing and those created in the same/forward direction to testing broken down by language pair, showing a mix of positive and negative differences in BLEU scores depending on test set creation direction.

tems participating in WMT-17 and WMT-18 news translation shared task broken down by language pair, where positive differences for systems indicate a higher human evaluation score when systems are tested in the reverse direction relative to the corresponding forward direction DA score.

As can be seen from the box plot in Figure 1 almost all reverse DA scores are higher than equivalent forward DA scores. This confirms the suspicion that absolute human evaluation results are in general higher when test data is created in the reverse direction to testing.

3.2 BLEU

Besides human evaluation, the performance of MT systems is often measured using automatic metrics, the most common of which remains to be the BLEU score (Papineni et al., 2002). Figure 2 shows a box plot of absolute differences in BLEU scores for systems (reverse BLEU − forward BLEU) participating in WMT news translation tasks from 2015 to 2018. Counter expectation there is a clear mix of positive and negative BLEU score differences for several language pairs.

Comparison of BLEU scores is not as straightforward as human evaluation however, and there are further consideration to be made before drawing conclusions from the mix of positive and negative absolute BLEU score differences described above. For example, the fact that splitting the test set into forward and reverse directions creates two test sets comprised of distinct sentences is likely to impact how each distinct BLEU score should be interpreted, as BLEU is not a simple arithmetic average of sentence scores (like human evaluation DA scores).

3.3 Relative Differences

Besides absolute differences in BLEU scores for individual systems, we also consider how differences correspond to one another for pairs of systems competing in the same competition. For example, for an individual competition, the problems associated with test data creation are more problematic if they occur differently for different systems and less severe if they affect all systems in the same way, as system scores are mainly interpreted relative to one another.

The scatter plot in Figure 3 shows relative differences in BLEU scores when we change from forward to reverse test data for all pairs of systems participating in WMT-15 to WMT-18, as well as differences in human DA scores for systems participating in WMT-17 to WMT-18. The absence of systems in the upper-left and lower-right quadrants reassuringly shows that although extreme changes in BLEU and human scores do occur when test set creation direction is altered, the changes are at least somewhat systematic in the sense that when a difference in scores occurs (a drop or increase

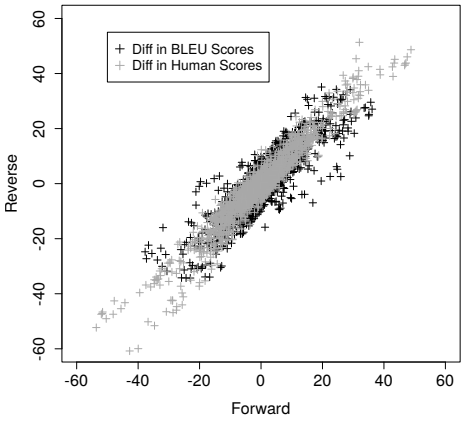

Figure 3: Differences in BLEU and human scores for pairs of systems participating in WMT-15 to WMT-18 and WMT-17 to WMT-18 respective news translation tasks for test sentences created in the reverse and forward directions.

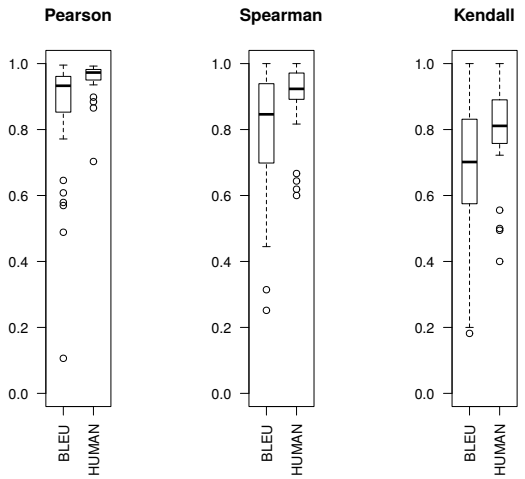

Figure 4: Pearson, Spearman and Kendall's τ correlation of forward and reverse BLEU and HUMAN scores of data for all available systems from WMT-15 – WMT-18 news translation task for BLEU and WMT-17–WMT-18 for human assessment.

when we change from forward to reverse test data), it occurs similarly for pairs of systems participating in the same competition. However, although there is a diagonal orientation in the plot, it still is somewhat worryingly broad and it remains possible that inclusion of reverse test data could bias BLEU and human scores in different ways for different types of systems.

3.4 System Rankings

Figure 4 shows Pearson, Spearman and Kendall's τ correlation of forward and reverse scores for systems participating individual competitions from WMT-15–WMT-18 terms of both BLEU and human evaluation. As can be seen, the correspondence between forward and reverse rank correlation of systems according to BLEU varies considerably across different evaluation test sets, from as low as a τ of 0.2 (tr-en newstest2018), where BLEU score rankings are extremely different depending on test data creation direction, up to a τ of 1.0, where rank correlation is identical (cs-en; fi-en newstest2017; fi-en; en-cs newstest2018).

In overall summary, our analysis of differences in both BLEU and human evaluation scores reveal differences in system rankings when tested on reverse and forward-created test data, differences substantial in some cases. Subsequently we have confirmed the validity of suspicions about lack of reliability of test data raised by Toral et al. (2018) and Läubli et al. (2018) caused by inclusion of reverse-created test data. However, as stated previously, neither reassessments of Hassan et al. (2018) ticked all boxes in terms of valid human evaluation methodologies and in the section that follows we therefore once again reassess the original evaluation.

4 Re-evaluation of Human Parity Claims

As described in detail in Section 2, past re-evaluations of human parity claims were hampered by sub-optimal test settings. In our re-evaluation, we firstly carry out statistical power analysis so that, in the case of encountering any ties between systems or indeed human and system, tests used to draw conclusions will have sufficient statistical power to avoid human-parity claims that in fact simply correspond to Type II error. Statistical power is of particular importance when considering document-level evaluation due to the fact that gathering ratings of documents as opposed to sentences requires substantially more annotation time and for this reason is highly likely to result in a reduction in the number of assessments collected in any evaluation. For example, Läubli et al. (2018) included as few as 55 documents in their re-evaluation of Hassan et al. (2018). Our concern about a potential substantial reduction in sample size in future document-level evaluations is well-founded therefore, especially considering standard segment-level MT human evaluations commonly include a sample of 1,500 segments. In the case of Läubli et al. (2018) this corresponds to an extreme reduction of

approximately 96% to the sample size. Since the very nature of the question being investigated involves a potential tie between human and machine, such a small sample size is a serious risk to the reliability of conclusions drawn simply due to its impact in terms of statistical power.

As a rough guide to what constitutes sufficient statistical power, we borrow the five-eighty convention from the behavioural sciences that provides a balance between Type I versus Type II error, where significance and power levels are set at 0.05 and 0.8 respectively (Cohen, 1988). Table 2 shows the statistical power, the probability of identifying a significant difference when one exists, of the statistical test applied in WMT evaluations, Wilcoxon rank-sum test, for a range of effect and sample sizes (n), where for the purpose of the test the appropriate effect size is the probability of the translations of system A being scored lower than those of system B. As shown in Table 2 for the usual sample size employed in WMT evaluations, 1,500, statistical power even for closely performing systems, where the effect size, the probability of the translations of system A being scored lower than those of system B, is 0.47, statistical power remains above 0.8. For such pairs of systems, however, if we were to employ the smaller sample size of 55 documents, as in Läubli et al. (2018), the power of the test to identify a significant difference falls to as low as 0.081, approaching one tenth of acceptable statistical power levels.[2]

A good compromise between fully document-level evaluation, where only ratings of documents are collected, and fully segment-level evaluations, in which segments are presented to human judges in isolation of the document, is collection of ratings of segments with the wider document context available to the human assessor and have the segments evaluated in their original order. In this way, a sufficient sample size can still be achieved to ensure appropriate levels of statistical power with the added aim of human judges being able to take into account the quality of translations within the wider document context.[3]

We therefore plan our re-evaluation as follows: (i) collect segment ratings for documents produced by a single system within the correct document context; (ii) aim to collect direct assessments of a sufficient number of translations exceeding the minimum acceptable sample size in terms of power analysis, approximately 385 distinct translations; (iii) use n, the number of distinct translations as opposed to repeat human assessments as the sample size; (iv) employ Direct Assessment, the most up to date technology for this purpose and that employed by WMT for the official results since 2017, a method shown to produce highly repeatable results; (v) only employ forward-created test data; (vi) only draw conclusions specific to Chinese to English translation and news domain; (vii) produce clusters with a standard significance test, Wilcoxon rank-sum test.

4.1 Re-evaluation Results

Direct Assessment (DA) HITs were set up and run as in WMT human evaluations on Mechanical Turk but with the distinction of segments being evaluated in the correct order in which they appeared in a document, comprising an initial set of results, which we refer to as segment rating + document context (SR+DC). In addition to the segment rating, workers were additionally shown entire documents and asked to rate them, providing a secondary set of results for comparison purposes. We refer to these fully document-level results as document rating + document context (DR+DC) configuration. As is usual in DA evaluations, translations were rated in a 0–100 scale and quality control was applied.

131 workers participated producing a total of 13,214 assessments of translations, of which 6,606 (49.99%) were from workers who passed DA's quality control checks. Table 3 shows results of our re-evaluation[4] of the top systems originally included in Hassan et al. (2018), where REF-HT is the original set of human translations produced by Hassan et al. (2018) themselves and against which human-parity of MT was claimed, while REF-PE is machine translated outputs that have been post-edited by humans, and Combo-6 is the best-performing system in Hassan et al. (2018).

Results when segments are rated by human judges within the correct document context (Segment Rating + Document Context) show that the DA score achieved by the human reference translation, REF-HT, is significantly higher than both

[2] In Läubli et al. (2018) the Sign test was used as opposed to Wilcoxon rank sum and has similar statistical power for such an effect size.

[3] This approach is not that of Toral et al. (2018), where document context was only available in for the source input document as opposed to MT output document.

[4] All evaluation data is publicly available at https://www.scss.tcd.ie/~ygraham/emnlp2020-translationese

	effect size																
n	0.330	0.340	0.350	0.360	0.370	0.380	0.390	0.400	0.410	0.420	0.430	0.440	0.450	0.460	0.470	0.480	0.490
55	**0.886**	**0.842**	0.788	0.725	0.659	0.586	0.512	0.438	0.367	0.300	0.243	0.188	0.144	0.111	0.081	0.066	0.056
330	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**0.995**	**0.982**	**0.947**	**0.878**	0.763	0.604	0.427	0.265	0.144	0.073
385	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.998**	**0.992**	**0.971**	**0.924**	**0.824**	0.672	0.485	0.302	0.159	0.077
440	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**0.997**	**0.986**	**0.951**	**0.870**	0.730	0.538	0.338	0.176	0.081
1485	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.997**	**0.965**	**0.809**	0.471	0.156
1540	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.998**	**0.971**	**0.821**	0.485	0.161
1595	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.998**	**0.975**	**0.838**	0.499	0.164

Table 2: Statistical Power of two-sided Wilcoxon Rank Sum Test for a range of sample and effect sizes; power $\geq$ 0.8 highlighted in bold.

Segment Rating + Document Context				
Ave.	Ave. z	n	N	System
80.3	0.143*	902	1811	REF-HT
76.6	0.038	904	1646	REF-PE
76.5	0.036	863	1805	Combo-6

Document Rating + Document Context				
Ave.	Ave. z	n	N	System
78.9	0.184	114	216	REF-HT
77.5	0.090	107	218	REF-PE
76.0	0.050	106	238	Combo-6

Table 3: Re-evaluation of human-parity-claimed Chinese to English system of Hassan et al. (2018); * denotes system that significantly outperforms all lower ranked systems according to a two-sided Wilcoxon rank-sum test $p < 0.05$

REF-PE and Combo-6, agreeing with results of both Läubli et al. (2018) and Toral et al. (2018). Since this approach has a large enough sample size to ensure sufficient statistical power, the tie between REF-PE and Combo-6 is a legitimate one however. Although this tie does indeed indicate high performance of Combo-6, since REF-PE is in fact post-edited MT output however, this tie does not provide legitimate evidence to support a human-parity claim.

Although we already know from the power analysis carried out for planning the current evaluation that fully document-level evaluations in which human assessors are required to rate documents (as opposed to segments) will encounter problems in terms of sufficient statistical power when ties occur, we nonetheless run this kind of evaluation for demonstration purposes. Document Rating + Document Context results in Table 3 do indeed produce what appears to be a statistical tie between the three sets of outputs as no "system" significantly outperforms all lower ranking ones. However, a conclusion of human parity cannot legitimately be claimed from this tie due to the low statistical power of the test caused by the small sample of documents that were rated. Ties in this case do not indicate human-parity but simply that the test is too weak to identify significant differences.

In summary, similar to Toral et al. (2018) and Läubli et al. (2018), our results show evidence that the original system, Combo-6, was outperformed by human translation. It should be noted however that from our results it cannot be inferred that machine translation in general has not yet reached human performance but simply that the system that originally claimed human-parity in fact did not achieve it, as tested on WMT-17 newstask data.

5 Conclusion

We explore issues relating to the reliability of machine translation evaluations. Firstly, we provide a detailed analysis of how the presence of translationese phenomena can adversely affect machine translation results. In terms of the legitimacy of machine translation evaluation results, our analysis provides sufficient evidence that translationese is a problem for evaluation of systems, in particular in terms of comparison of system performance with automatic metrics such as BLEU. This results in our first recommendation in future MT evaluations to *avoid the use of source side test data that was created via human translation from another language*. We provided guidance in relation to sample size and statistical power to help planning future human evaluations of MT, particularly relevant to document-level human-parity investigations. This guidance will help to avoid false conclusions due to the application of low powered statistical tests.

Acknowledgments

This study was supported by the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Trinity College Dublin funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund, and it has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825299 (Gourmet). We would also like to thank the anonymous reviewers for their feedback.

References

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology: In Honour of John Sinclair*, Netherlands. John Benjamins Publishing Company.

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.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

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

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

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. 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*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics. Revised August 2010.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical power analysis for the social sciences*. Hillsdale, NJ: Erlbaum.

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

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 (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

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

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in Machine Translation Evaluation. *arXiv e-prints*, page arXiv:1906.09833.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Gennadi Lambersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38:4.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Neural Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *EMNLP 2018*, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *CoRR*, abs/1808.10432.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.