Further Investigation into Reference Bias in Monolingual Evaluation of Machine Translation

Qingsong Ma[*†] **Yvette Graham**[†] **Timothy Baldwin**[‡] **Qun Liu**[†]

[*]Institute of Computing Technology
Chinese Academy of Sciences
maqingsong@ict.ac.cn

[†]ADAPT Centre
Dublin City University
firstname.surname@dcu.ie

[‡]Computing and Info Systems
University of Melbourne
tb@ldwin.net

Abstract

Monolingual evaluation of Machine Translation (MT) aims to simplify human assessment by requiring assessors to compare the meaning of the MT output with a reference translation, opening up the task to a much larger pool of genuinely qualified evaluators. Monolingual evaluation runs the risk, however, of bias in favour of MT systems that happen to produce translations superficially similar to the reference and, consistent with this intuition, previous investigations have concluded monolingual assessment to be strongly biased in this respect. On re-examination of past analyses, we identify a series of potential analytical errors that force some important questions to be raised about the reliability of past conclusions, however. We subsequently carry out further investigation into reference bias via direct human assessment of MT adequacy via quality controlled crowd-sourcing. Contrary to both intuition and past conclusions, results show no significant evidence of reference bias in monolingual evaluation of MT.

1 Introduction

Despite it being known for some time now that automatic metrics, such as BLEU (Papineni et al., 2002), provide a less than perfect substitute for human assessment (Callison-Burch et al., 2006), evaluation in MT more often than not still comprises BLEU scores. Besides increased time and resources required by the alternative, human evaluation of systems, human assessment of MT faces additional challenges, in particular the fact that human assessors of translation quality tend to be highly inconsistent. In recent Conference on Machine Translation (WMT) shared tasks, for example, manual evaluators complete a relative ranking (RR) of the output of five alternate MT systems, where they must rank the quality of competing translations from best to worst. Within this set-up, when presented with the same pair of MT output translations, human assessors often disagree with one another's preference, and even their own previous judgment about which translation is better (Callison-Burch et al., 2007; Bojar et al., 2016). Low levels of inter-annotator agreement in human evaluation of MT not only cause problems with respect to the reliability of MT system evaluations, but unfortunately have an additional knock-on effect with respect to the meta-evaluation of metrics, in providing an unstable gold standard. As such, provision of a fair and reliable human evaluation of MT remains a high priority for empirical evaluation.

Direct assessment (DA) (Graham et al., 2013, 2014, 2016) is a relatively new human evaluation approach that overcomes previous challenges with respect to lack of reliability of human judges. DA collects assessments of translations separately in the form of both fluency and adequacy on a 0–100 rating scale, and, by combination of repeat judgments for translations, produces scores that have been shown to be highly reliable in self-replication experiments (Graham et al., 2015). The main component of DA used to provide a primary ranking of systems is *adequacy*, where the MT output is assessed via a monolingual similarity of meaning assessment. A reference translation is displayed to the human assessor (rendered in gray) and below it the MT output (in black), with the human judge asked to state the degree to which they agree that *The black text adequately expresses the meaning of the gray text in English.*[1] The motivation behind

[1]Instructions are translated into a given target language.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2466–2475
Copenhagen, Denmark, September 7–11, 2017.

constructing DA as a monolingual MT evaluation are as follows:

- Monolingual assessment of MT opens up the annotation task to a larger pool of genuinely qualified human assessors;
- Crowd-sourced workers are unlikely to make use of information that is not entirely necessary for completing a given task; and are therefore unlikely to use the source language input if the reference is also displayed or to make use of the source input inconsistently;
- Displaying only the source without a reference greatly increases both the difficulty of the task and the time required to complete each annotation, which is too serious a trade-off when we wish to carry out human assessment on a very large scale;
- Varying levels of proficiency in the source language across different human assessors could contribute to inconsistency in bilingual MT evaluations.

Although DA has been shown to overcome the long-standing challenge of lack of reliability in human evaluation of MT, the possibility still exists that, although scores collected with DA have been shown to be almost perfectly reliable in self-replication experiments, both sets of scores, although consistent with each other, could in fact both be biased in the same way. Graham et al. (2013) include in the design of DA a number of criteria aimed at minimizing such bias: (i) assessment of individual translations in isolation from others to avoid a given system being scored unfairly low due to its translations being assessed more frequently alongside high quality translations (Bojar et al., 2011); (ii) elicit assessment scores via a Likert-style question without intermediate labeling, motivated by medical research showing patients' ratings of their own health to be highly dependent on the exact wording of descriptors (Seymour et al., 1985); (iii) accurate quality control by assessing the consistency of judges with reference only to their own rating distributions, to accurately remove inconsistent crowd-sourced data while avoiding removal of data that legitimately diverges from the scoring strategy of a given expert judge; and (iv) score standardization to avoid bias introduced by legitimate variations in scoring strategies.

Despite efforts to avoid bias in Graham et al. (2013), since DA is a monolingual evaluation of MT that operates via comparison of MT output with a reference translation, it is therefore still possible, while avoiding other sources of bias, that DA incurs reference bias where the level of superficial similarity of translations with reference translations results in an unfair gain, or indeed an unfair disadvantage for systems that yield translations that legitimately deviate from the surface form of reference translations. Following this intuition, Fomicheva and Specia (2016) carry out an investigation into bias in monolingual evaluation of MT and conclude that in a monolingual setting, human assessors of MT are strongly biased by the reference translation. In this paper, we provide further analysis of experiments originally provided in Fomicheva and Specia (2016), in addition to further investigation into the degree to which the intuition about reference bias can be supported.

2 Background

Fomicheva and Specia (2016) provide an investigation into reference bias in monolingual evaluation of MT. 100 Chinese to English MT output translations are assessed by 25 human judges on a five-point scale, in the form of their response (*None*, *Little*, *Much*, *Most*, or *All*) to the following question: *how much of the meaning of the human translation is also expressed in the machine translation?*. Precisely the same 100 translations were assessed by all 25 judges. Human judges were divided into five groups of five: Group 1 (G_1) was shown the source language input and the MT output only and carried out a bilingual assessment, while Groups 2–5 (G_2–G_5) were not shown the source input but instead compared the MT output to a human-generated reference translation. A distinct set of reference translations was assigned to each group G_2–G_5. Inter-annotator agreement (IAA) was measured for pairs of judges as follows (the total number of judge pairs resulting from each setting is provided in parentheses):

- SOURCE: a given pair of judges assessed translations in a bilingual setting (all possible pairs within G_1 = $\binom{5}{2} = 10$ pairs);
- SAME: a given pair of judges assessed translations in a monolingual setting by comparison with precisely the *same reference translation* (the sum of all possible pairs result-

DIFF	SAME	SOURCE
0.163 ± 0.01	0.197 ±0.01	0.190 ± 0.02

Table 1: Average Kappa coefficients and 99% confidence intervals reported in Fomicheva and Specia (2016)

ing from each individual group G_2–G_5 = $\binom{5}{2} + \binom{5}{2} + \binom{5}{2} + \binom{5}{2} = 40$ pairs);

- DIFF: a given pair of judges assessed translations in a monolingual setting by comparison with a *distinct reference translation* (cross product of judges belonging to the four groups G_2–G_5 = $G_2 \times G_3 \times G_4 \times G_5$ = 150 pairs).

Reference bias is investigated by comparison of levels of IAA, via Cohen's Kappa (κ) and weighted Kappa coefficients. The hypothesis, although not explicitly stated, is that if agreement of human assessors of MT in SAME is *higher* than that of assessors in DIFF, then the likely cause is reference bias in human assessment scores. Agreement in terms of Cohen's Kappa reported in Fomicheva and Specia (2016) are reproduced here in Table 1, where a small increase of 0.034 in average Kappa is shown for pairs of human assessors in SAME over that of DIFF. To avoid drawing conclusions from a difference that is likely to have occurred simply by chance, confidence intervals (CIs) are provided and the non-overlapping CIs for SAME and DIFF shown in Table 1 provide the basis for the conclusion that IAA is significantly higher for SAME compared to DIFF and subsequently that monolingual evaluation of MT is strongly biased by the reference translation. On examination of the analysis that led to the conclusion of strong reference bias, we unfortunately discover a series of methodological issues with respect to confidence interval estimation, however, that raise doubt about the reliability of this conclusion.[2]

A clear indication of the precise approach to CI estimation attempted in Fomicheva and Specia (2016) is unfortunately not explicitly stated but out of the range of methods that exist the approach that is applied most resembles bootstrap resampling. Conventionally speaking, bootstrap resampling can be applied to CI estimation of a point estimate for a *sample*, D, of size N, by simulating the variance in the *population* sampling distribution (Efron and Tibshirani, 1993). A standard method of estimating CIs via bootstrap resampling is to generate a bootstrap distribution for the statistic of interest made up of M repeat computations of it, each time drawing a random sample of size N from D with replacement. Although most similar to bootstrap resampling, the application in Fomicheva and Specia (2016) to CI estimation of Kappa coefficients diverges in some important ways from a standard application, however. We therefore provide a comparison of the analysis drawn in Fomicheva and Specia (2016) with a standard bootstrap implementation.

Figure 1(a) shows SAME and DIFF bootstrap distributions, reproduced from code released with Fomicheva and Specia (2016), originally yielding non-overlapping CIs that led to the conclusion of strong reference bias.[3] Although the level of statistical significance is reported to be 99%, CIs in Figure 1(a) show that the proportion of each bootstrap distribution was substantially underestimated leading to overly narrow CI limits for both SAME and DIFF. In contrast, Figure 1(b) shows CIs resulting from an accurately computed proportion of 95% of the same bootstrap distribution, where even at the lower level of 95% significance (as opposed to 99%) CIs for SAME and DIFF now overlap, reversing the conclusion of strong reference bias.

In addition, CI estimation diverges from bootstrap resampling with respect to the number of bootstrap samples employed. Since there are a total of N^N possible distinct bootstrap samples for a given sample D (taking order into account), in a conventional bootstrap implementation a Monte Carlo approximation of size M is employed, and the larger M is, the closer the distribution approaches the true bootstrap distribution (Chernick and LaBudde, 2014). In Fomicheva and Specia (2016), CIs are computed via only 50 bootstrap samples, however.[4] Figure 1(c) shows the change in location of CIs for a typical M =1,000, as opposed to M =50 (Figure 1(b)).

[2]We provide a re-analysis of experiment data specifically with respect to Cohen's Kappa. All errors outlined for Cohen's Kappa also lead to the same inaccuracies for weighted Kappa in Fomicheva and Specia (2016), however.

[3]CIs correspond closely to those of the original (Table 1) but differ by a tiny amount due to the randomness involved in regeneration from the code.

[4]We note that M is described as 100 in the publication, but 50 in the released code. Our question raised about the methodology also stands for M =100.

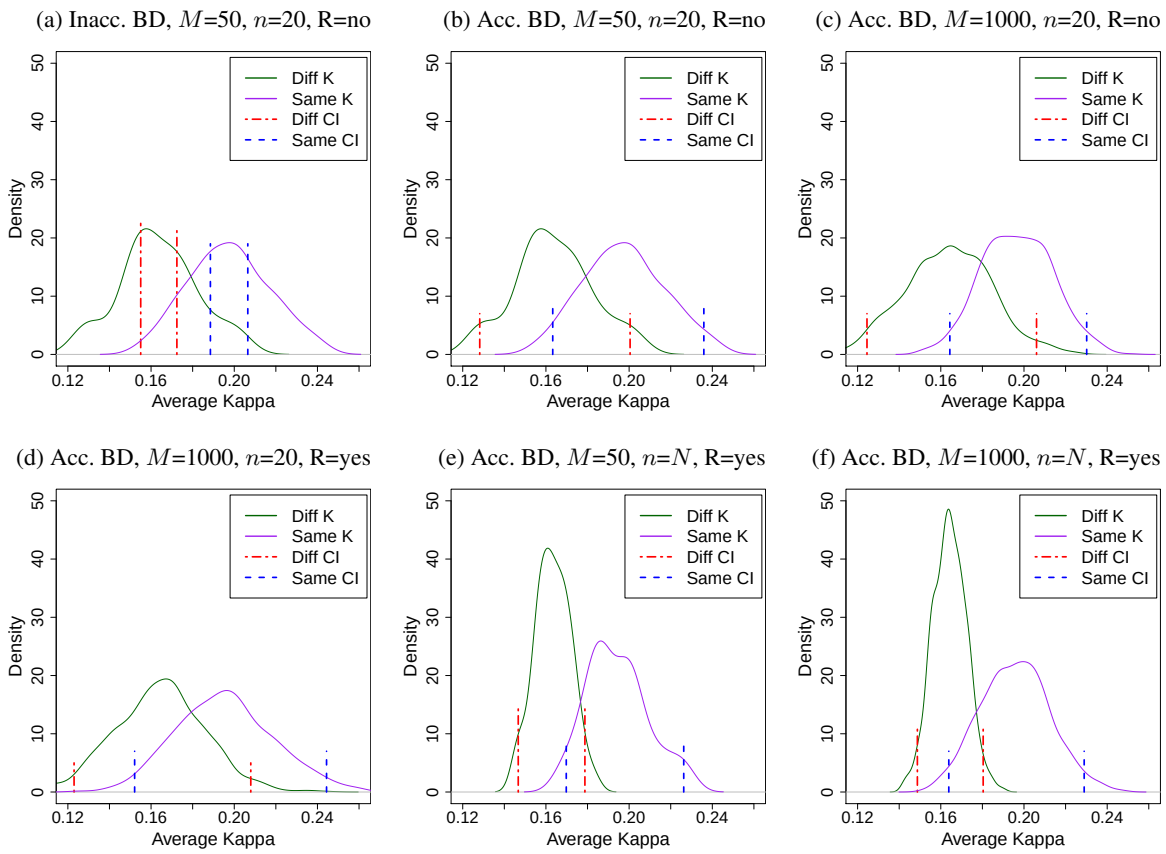


Figure 1: (a) Original bootstrap distribution (BD) and confidence intervals (CI) for average Kappa coefficients when human annotators employ the same reference translation (Same K) or a different reference translation (Diff K) in Fomicheva and Specia (2016) ("Inacc. BD"=inaccurate BD proportion; "Acc. BD"=accurate BD proportion; "M"=number of bootstrap samples; "n"=bootstrap sample size; "R=yes": sampled with replacement; "R=no": sampled without replacement); (b) is (a) with accurate BD proportion; (c) is (b) with conventional M; (d) is (c) with R=yes; (f) is (d) with N=n (N is the full sample size); (e) is (f) with M=50; (f) corresponds to correct BD with all CI errors corrected.

Furthermore, bootstrap distributions in Fomicheva and Specia (2016) are computed by random sampling *without* replacement, and the size of each bootstrap does not equal the original sample size N.[5] Figure 1(d) shows bootstrap distributions of Figure 1(c) when the sampling without replacement error is corrected, and Figure 1(f) shows bootstrap distributions of Figure 1(d) when the sample size error is corrected.

In summary, Figure 1(f) shows all errors with respect to CI estimate in Fomicheva and Specia (2016) corrected, and subsequently CIs for a standard implementation of bootstrap, which can be contrasted to those that led to the original conclusion of strong reference bias in Figure 1(a). CIs in Figure 1(f) for SAME and DIFF now overlap revealing that experiments in Fomicheva and Specia (2016), thus far do not show any evidence of reference bias.

2.1 Measures of Central Tendency

Even if the correct implementation of bootstrap resampling, shown in Figure 1(f), had shown non-overlapping confidence intervals, it would still unfortunately not have been appropriate to draw a conclusion from this of reference bias, however, due to the fact that significant differences are not investigated for the statistic of interest, the Kappa coefficient, but only for a measure of central tendency of two Kappa coefficient distributions, the *average* Kappa of each Kappa distribution. One reason for avoiding a comparison based on significant differences in average Kappa, as opposed to the Kappa point estimates themselves, is that it is

[5] A variant of bootstrap does exist where N is intentionally lowered to appropriately reduce the variance estimate but is only applicable when that of standard bootstrap is known to be over-estimated (Chernick and LaBudde, 2014).

possible for the average of two distributions to be equal, or indeed have a small but non-significant difference, while the underlying distributions differing considerably in several other respects.

Figure 2 shows Kappa coefficient distributions for all pairs of judges in SAME (40 pairs), DIFF (150 pairs) and SOURCE (10 pairs), revealing all distributions to have very similar Kappa coefficient distributions, with the one exception arising for SOURCE, where two of the human annotator pairs had an unusually high agreement level.[6]

A more informative comparison about levels of agreement in SAME and DIFF examines significant differences in Kappa point estimates, as opposed to comparison based on a measure of central tendency. For this reason, despite there being no significant difference in *average Kappa* for SAME and DIFF, we also examine the *proportion of Kappa point estimates of judge pairs in* SAME *that are significantly different from agreement levels of judge pairs in* DIFF, which will provide genuine insight into differences in levels of agreement between the two groups.

Table 2 shows proportions of all judge pairs with significant differences in Kappa point estimates (non-overlapping confidence intervals) for each combination of settings (Revelle, 2014).[7] The number of significant differences in Kappa point estimates for pairs of judges in SAME and DIFF is only 13%, or, in other words, 87% of judge pairs across SAME and DIFF have no significant difference in agreement levels. Table 2 also includes proportions of significant differences for Kappa point estimates resulting from judges belonging to a single setting (significance testing all Kappa of SAME with respect to all *other* Kappa belonging to SAME, for example), revealing that the proportion of significant differences within SAME (12%) to be very similar to that of SAME × DIFF (13%), and similarly for DIFF (12%), with only a single percentage point difference in both cases in proportions of significant differences. Subsequently, even after correcting the measure of central tendency error in Fomicheva and Specia (2016), evidence of reference bias can still not be concluded.

[6]The difference in distributions for SOURCE is exaggerated to some degree due to the total number of annotator pairs in SOURCE being substantially lower than the other two settings (only 10 pairs).

[7]Our re-analysis code is available at https://github.com/qingsongma/percentage-refBias

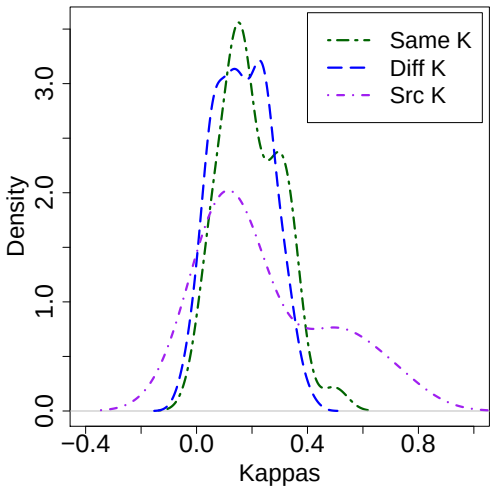


Figure 2: Distribution of Kappa coefficients for translations assessed with the same reference translation ("Same K"), different reference translations ("Diff K") and source sentences ("Src K") (Fomicheva and Specia (2016) data set).

	SOURCE	SAME	DIFF
SOURCE	47% (45)	29% (400)	27% (1,500)
SAME	–	12% (780)	13% (6,000)
DIFF	–	–	12% (11,175)

Table 2: Percentage of human annotator pairs in Fomicheva and Specia (2016) with significant differences in Kappa coefficients for pairs of annotators shown the same reference translation (SAME), different reference translations (DIFF) or the source language input only (SOURCE), total numbers of annotator comparisons in each case are provided within parentheses, numbers of annotator pairs was 10 for SOURCE, 40 for SAME and 150 for DIFF.

2.2 Differences in Ratings

The effect that reference bias may or may not have on actual 1–5 ratings attributed to translations, is again only reported in terms of a measure of central tendency, i.e. average ratings, in Fomicheva and Specia (2016). The average rating of each group shown a distinct reference translation is reported, showing distinct average scores for assessors employing a distinct set of reference translations. Due to the fact that each group had a distinct average rating, the conclusion is drawn that *MT quality is perceived differently depending on the human translation used as gold-standard*. It is however, entirely possible that, the difference in average ratings is in part or even fully caused by

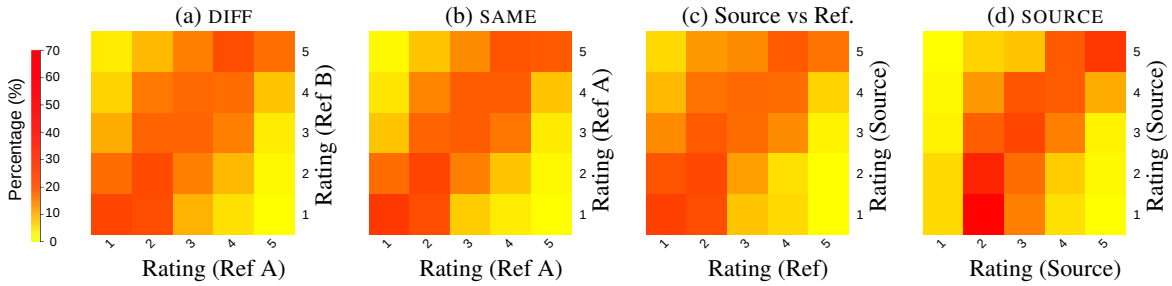


Figure 4: Proportions of 1–5 ratings (1=lowest; 5=highest) for translations when human assessors are shown different reference translations (DIFF), the same reference translation (SAME), the source input versus a reference translation (Source vs Ref.) or the source input (SOURCE) for data in Fomicheva and Specia (2016).

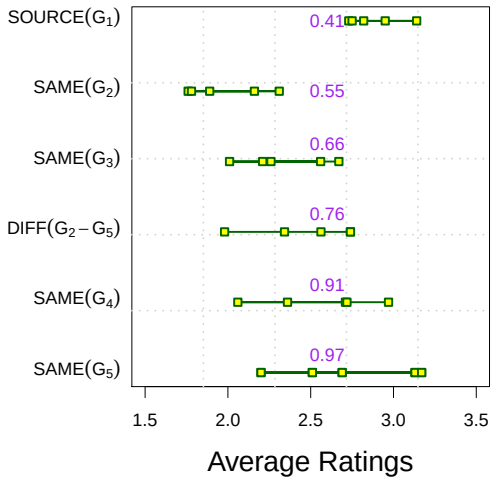


Figure 3: Average rating of human assessors shown the source input (SOURCE), the same reference translation (SAME), or a distinct reference translation (DIFF); range of average ratings provided adjacent to each setting.

the known lack of consistency across human annotators in general.

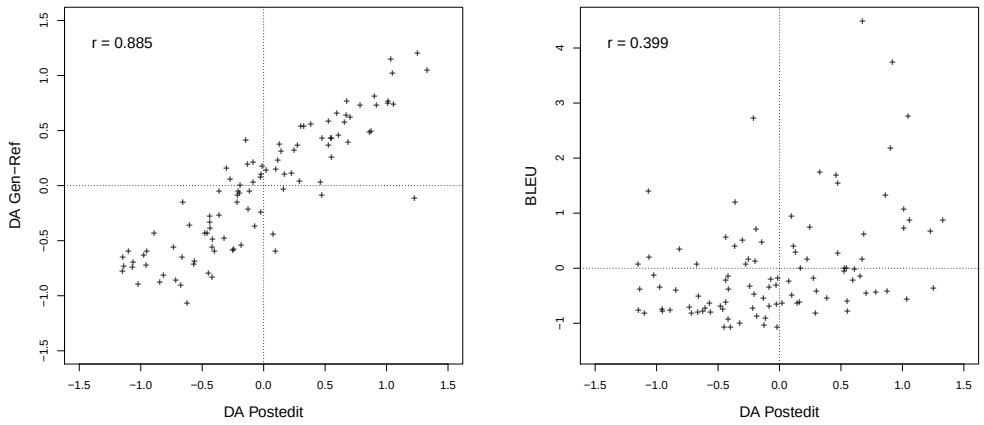
Quite a substantial leap is made therefore between the difference in average ratings and the cause of that difference. To investigate this further, we reproduce the average ratings for assessors shown a distinct reference translation, each represented by a green square along the line labeled "DIFF(G_2–G_5)" in Figure 3, where the overall range in average ratings is 0.76. The extremity of this range is better put into context by comparison with the average rating of human assessors shown the same reference translation, each labeled SAME in Figure 3, where the range of average ratings attributed to human assessors shown *the same reference* can be as large as 0.97 (G5). Thus, it *cannot* be concluded from a difference in average ratings for annotators shown distinct reference translations that the cause of this difference is the reference translation.

However, comparison of ratings based only on averages, again hides detail that an analysis could otherwise benefit from. We therefore examine the distribution of individual ratings attributed to translations, and how well ratings for the same translation correspond when pairs of annotators employ the same or distinct reference translation (or indeed the source input) in Figure 4.[8] The rating pattern in Figure 4 (a) of judge pairs employing a distinct reference translation compared to those in Figure 4 (b), where assessors employ the same reference translation, shows agreement at the level of individual ratings to be almost indistinguishable, showing no evidence of reference bias.

3 Alternate Reference Bias Investigation

Although we can now say that experiments in Fomicheva and Specia (2016) showed no evidence of reference bias, a further issue lies in the fact that low IAA was incurred throughout the study, and low IAA unfortunately provides no assurance with respect to the reliability of conclusions, even when corrected for analytical errors. In addition, the fact that IAA was itself the measure by which bias was investigated is also likely to exacerbate any problems with respect to reliability of conclusions. We therefore provide our own additional investigation into reference bias in monolingual evaluation of MT. Instead of investigating via IAA, we explore the degree to which unfairly high or low ratings might be assigned to translations with respect to

[8]The sum of percentages in a given row equals 100% in each heat map.



(a) (b)

Figure 5: (a) Scatter plot of direct assessment (DA) scores for 100 Chinese to English translations carried out by comparison with a generic reference translation (DA Gen-Ref) or DA with the reference replaced by a human post-edit of the MT output (DA Postedit); (b) sentence-level (smoothed) BLEU scores for the same translations also plotted against DA POST-EDIT; translations and references of (a) and (b) data set of Fomicheva and Specia (2016); post-edits provided by professional translators with access to the source and MT output only. BLEU and DA scores are standardized for ease of comparison in all plots.

surface similarity or dissimilarity with the reference translation.

Reference-similarity bias is the attribution of unfairly high scores to translations due to high surface-similarity with the reference translation even though the translation is not high quality. A converse kind of reference bias can also occur, which we call *reference-dissimilarity bias*, where unfairly low scores are attributed to translations that are superficially dissimilar to the reference translation but are in fact high quality translations. The challenge in investigating reference bias lies in the ability to accurately distinguish between translations that receive *unfair* scores due to surface-similarity or dissimilarity from those that achieved a *fair* score due to the translation being genuinely high or low quality.

To separate genuine high quality translations from those that score *unfairly* high, we carry out two separate assessments of the same set of translations. Firstly, we carry out a standard monolingual MT evaluation that employs a generic reference translation (GEN-REF setting), the scores that potentially encounter reference bias. Secondly, we carry out an additional human evaluation of the same translations, where, instead of the generic reference, the human assessor compares the MT output with a human post-edit of it (POST-EDIT setting). The latter human assessment is highly unlikely to encounter any form of reference bias because the assessment employs a post-edit of the MT output, which itself will only differ the MT output with respect to the parts of it that are genuinely incorrect. Translations encountering reference-similarity bias can then be identified by a high GEN-REF score combined with a low POST-EDIT score, and vice-versa for reference-dissimilarity, a low GEN-REF score combined with a high POST-EDIT score.

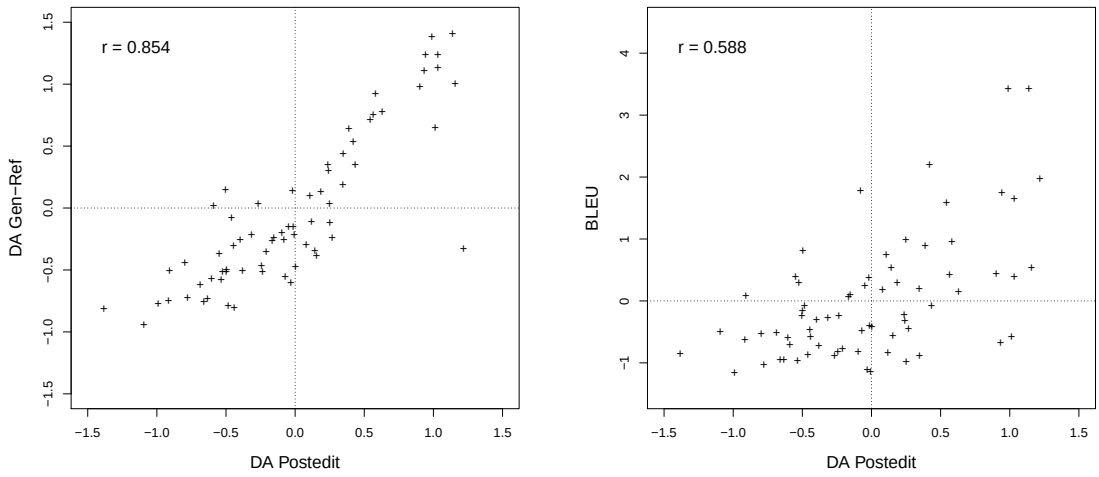
3.1 Reference Bias Experiments

Experiments were carried out using the original 100 Chinese to English translations released by Fomicheva and Specia (2016), in addition to 70 English to Spanish MT translations (WMT-13 Quality Estimation Task 1.1).[9] Professional translators, entirely blind to the purpose of the study, were employed to post-edit the MT outputs used in the POST-EDIT setting, and were shown the source input document and the MT output document only (no reference translations).[10]

Once post-edits had been created, DA was employed in two separate runs on Amazon Mechani-

[9] A single generic reference translation was chosen at random from the Chinese to English data set; only a single reference is available for each translation in the English to Spanish data set.

[10] Post-editors were paid at the standard rate.



(a) (b)

Figure 6: (a) Scatter plot of direct assessment (DA) scores for 70 English to Spanish translations carried out by comparison with a generic reference translation (DA Gen-Ref) or DA with the reference replaced by a human post-edit of the MT output (DA Postedit); (b) sentence-level (smoothed) BLEU scores for the same translations also plotted against DA POST-EDIT; translations and generic references for (a) and (b) WMT-13 Quality Estimation Task 1.1 (Bojar et al., 2013) data set; post-edits provided by professional translators with access to the source and MT output only. DA and BLEU scores are standardized for ease of comparison in all plots.

cal Turk,[11] once for GEN-REF and once for POST-EDIT. Besides employing distinct reference translations in the assessment, all other set-up criteria were identical for both evaluation settings, including the conventional segment-level DA setting, where a minimum of 15 human assessments are combined into a mean DA score for a given translation, after strict quality control measures and score standardization have been applied.

3.2 Results and Discussion

Figure 5(a) shows a scatter-plot of DA scores attributed to translations for GEN-REF compared to POST-EDIT in the Chinese to English experiment. Translations that encounter reference-dissimilarity bias are expected to appear in the lower-right quadrant of Figure 5(a), receiving an unfairly low GEN-REF score combined with a high POST-EDIT score. As can be seen from Figure 5(a) only a very small number of translations fall into this quadrant, all of which are very closely located to adjacent upper-right and lower-left quadrants. A single translation in Figure 5(a) is an outlier in this respect, receiving a high POST-EDIT score in combination with a lower than average GEN-REF score, possibly indicating reference bias. On closer inspection, however, the score combination is in fact the result of a mistake in the reference translation. Although the low GEN-REF score was the result of an error in the reference translation, a single translation having this score combination is not sufficient evidence to conclude strong reference bias. In future work we would like to investigate the frequency of erroneous reference translations in existing MT test sets, although we expect them to be few, accurate statistics would provide a better indication of the degree to which they could negatively impact the accuracy of DA evaluations.

Figure 5(a) is also void of evidence of reference-similarity bias, as only a small number of translations lie in the upper-left quadrant and are all very close to the origin and/or adjacent quadrants.

Contrasting Figure 5(a), the correspondence of GEN-REF scores to POST-EDIT scores, with Figure 5(b), the correspondence of known reference-biased BLEU scores, in contrast a large number of BLEU scores for translations do encounter reference bias, as seen by the spread of translations appearing across both the bottom-right and upper-left quadrants.

[11] https://www.mturk.com

Similarly for English to Spanish, the correspondence between GEN-REF and POST-EDIT scores for translations are shown in Figure 6(a), where, again, only a small number of translations appear in the bottom-right and upper-left quadrants, all lying very close to adjacent quadrants, again, showing no significant indication of reference bias. A single translation appears to break the trend again, however, receiving a low GEN-REF score combined with a high POST-EDIT score, located in the lower-right quadrant of Figure 6(a). On closer inspection, the low GEN-REF score is the result of something unexpected, as the MT output is in fact an accurate translation while at the same time the generic reference is also correct, but unusually the meaning of the two diverge from each other.[12] Again, a single translation receiving this score combination is not sufficient evidence to conclude reference bias to be a significant problem for monolingual evaluation. The lack of reference bias in Figure 6(a) can again be contrasted to known reference-biased BLEU scores in Figure 6(b) for English to Spanish.

4 Conclusions

In this paper, we provided an investigation into reference bias in monolingual evaluation of MT. Our review of past investigations reveals potential analytical errors and raises questions about the reliability of past conclusions of strong reference bias. This motivates our further investigation for Chinese to English and English to Spanish MT employing direct human assessment in a monolingual MT evaluation setting. Results showed no significant evidence of reference bias, contrary to prior reports and intuition.

Acknowledgments

This project has received funding from NSFC Grant No. 61379086 and the European Union Horizon 2020 research and innovation programme under grant agreement 645452 (QT21) and Science Foundation Ireland in the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund.

[12] Source: *A straightforward man*; MT: *Un hombre sencillo*; Reference: *Un hombre sincero*

References

Ondrej Bojar, Miloš Ercegovčevic, Martin Popel, and Omar F. Zaidan. 2011. A grain of salt for the WMT manual evaluation. In *Proceedings of the 6th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, pages 1–11.

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*. Association for Computational Linguistics, Sofia, Bulgaria, pages 1–44.

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*. Association for Computational Linguistics, Berlin, Germany, pages 131–198.

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*. Association for Computational Linguistics, Prague, Czech Republic, pages 136–158.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Trento, Italy, pages 249–256.

Michael R Chernick and Robert A LaBudde. 2014. *An introduction to bootstrap methods with applications to R*. John Wiley & Sons.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York City, NY.

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*. Association for Computational Linguistics, Berlin, Germany, pages 77–82.

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. Association for Computational Linguistics, Sofia, Bulgaria, pages 33–41.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, pages 443–451.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone? *Natural Language Engineering* pages 1–28. https://doi.org/10.1017/S1351324915000339.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1183–1191.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, PA, pages 311–318.

William Revelle. 2014. psych: Procedures for personality and psychological research. *Northwestern University, Evanston. R package version* 1(1).

Robin A. Seymour, Judy. M. Simpson, J. Ed Charlton, and Michael E. Phillips. 1985. An evaluation of length and end-phrase of visual analogue scales in dental pain. *Pain* 21:177–185.