Is all that Glitters in MT Quality Estimation really Gold Standard?

Yvette Graham[*] **Timothy Baldwin**[†] **Meghan Dowling**[*] **Maria Eskevich**[‡]
Teresa Lynn[*] **Lamia Tounsi**[*]

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

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

[‡]Centre for Language Studies
Radboud University
m.eskevich@let.ru.nl

Abstract

Human-targeted metrics provide a compromise between human evaluation of machine translation, where high inter-annotator agreement is difficult to achieve, and fully automatic metrics, such as BLEU or TER, that lack the validity of human assessment. Human-targeted translation edit rate (HTER) is by far the most widely employed human-targeted metric in machine translation, commonly employed, for example, as a gold standard in evaluation of quality estimation. Original experiments justifying the design of HTER, as opposed to other possible formulations, were limited to a small sample of translations and a single language pair, however, and this motivates our re-evaluation of a range of human-targeted metrics on a substantially larger scale. Results show significantly stronger correlation with human judgment for HBLEU over HTER for two of the nine language pairs we include and no significant difference between correlations achieved by HTER and HBLEU for the remaining language pairs. Finally, we evaluate a range of quality estimation systems employing HTER and direct assessment (DA) of translation adequacy as gold labels, resulting in a divergence in system rankings, and propose employment of DA for future quality estimation evaluations.

1 Introduction

Although human evaluation of translation quality in theory provides the most meaningful assessment of machine translation (MT), achieving high levels of agreement between human assessors has proven challenging. For example, the annual Workshop on Statistical Machine Translation (WMT) provides large-scale human evaluation of systems, and reports inter-annotator agreement ranging from 0.260 (WMT-13) to 0.405 (WMT-15), with intra-annotator agreement not faring much better, from 0.407 (WMT-12) to 0.595 (WMT-15) (Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2015).[1] Low agreement levels in human annotation cause challenges for tasks that require evaluation of MT output on the segment-level. For example, evaluation of MT quality estimation (QE) requires the comparison of system predictions of translation quality with segment-level human assessment. Automatic metrics, such as BLEU or TER, although fully repeatable, unfortunately lack the validity of human assessment and are well-known to suffer from bias in favour of translations that happen to be superficially similar to reference translations.

Part-automatic human-targeted metrics, however, are commonly employed as a substitute for human assessment on the segment-level, as they appear to provide a happy medium between fully automatic metrics, that lack the validity of human assessment, and human assessment that lacks reliability/reproducibility of automatic metrics. Human-targeted reference translations, each manually created by minimally post-editing the individual MT output translation to be assessed, remove the bias usually introduced by comparison with a generic reference. Subsequently an automatic component is applied to quantify the error now present between the MT output and its human-targeted reference.

[1] Agreement levels provided are average Kappa coefficients for all language pairs included in the main translation shared task.

Although human-targeted metrics are indisputably more valid than their generic-reference counterparts, the scores they produce are nonetheless still partly automatic. Given the vast number of possible methods of comparing a given MT output with its (human-targeted) reference translation, it is necessary to provide evidence that any given choice of human-targeted metric provides the best formulaic substitute for human assessment.

As with fully automatic metrics, human-targeted metrics are themselves evaluated by strength of correlation with human assessment. For example, when the most widely applied human-targeted metric, human-targeted translation edit rate (HTER) was first proposed, Snover et al. (2006) reported that "HTER is more highly correlated to human judgments than BLEU or HBLEU" (p. 230), and hence, since 2006, it is HTER, as opposed to HBLEU, that is employed in MT as a human assessment substitute.

2 Relevant Work

HTER is generally regarded as a valid substitute for human assessment. In MT QE, for example, HTER scores are used to evaluate systems in large-scale shared tasks (Bojar et al., 2015; Graham, 2015). If a metric such as HTER is to be relied upon as a substitute for human assessment, it is important that trust in the metric is well-placed to avoid inaccuracies in empirical evaluations. For instance, Bojar et al. (2013) and Graham (2015) make the assumption that HTER provides a valid representation of translation quality and subsequently employ HTER scores as a gold standard representation when evaluating QE systems, ultimately leading to rankings for competing systems. If HTER scores do not in fact provide a valid representation of translation quality, however, system rankings are likely to be incorrect. On review of experiments that originally led to trust in HTER as a substitute for human assessment, a possible disparity emerges between the degree of trust placed in HTER and the limitations of original experiments. These limitations include:

- A sample of translations produced by two distinct MT systems, one of known *low-performance* and one known *high-performance* system;
- A single language pair (Arabic to English);
- A maximum of two human annotators per translation;
- Employment of human-targeted reference translations created by human post-editors coached to specifically minimize HTER scores as opposed to other possible formulations.

The degree to which general conclusions can be drawn from experiments should reflect the extent to which experiments were carried out, and the broad conclusions drawn that HTER has a strong correlation with human assessment is not ideal considering experiments were limited to a small number of translations produced by two MT systems and for a single language pair. Furthermore, translations sampled from systems that operate at the two extremes of performance, a known high quality system and a poorly performing one, risks exaggeration of correlation with human assessment.

In addition, it is possible that issues in relation to levels of inter-annotator agreement, common in other human evaluations, were also present in the human evaluation used in HTER experiments, as Snover et al. (2006) report that "... the four-reference variants of TER and HTER correlate with human judgments as well as – or better than – a second human judgment does" (p. 223). The fact that HTER scores correlated with the quality assessments of one human assessor more than those of another, highlights the likelihood that agreement between human annotators was low, although no precise agreement levels are reported. If, for example, the correlation of scores produced by a (part-automatic) metric, such as HTER, and human assessment scores provided by a given human assessor, Annotator$_a$, is stronger than the correlation between Annotator$_a$ and a second human annotator, Annotator$_b$, it follows that this latter correlation, between human assessments of the two distinct annotators cannot be all that high. Choosing to trust the annotations of Annotator$_a$ simply on the basis that those annotations yield a more favorable correlation with HTER is not justified. The fact that part-automatic metric scores correlate better with one human annotator than another is not evidence of how well the metric is performing, merely that the human evaluation employed is somewhat unstable.

	into English					out of English		
	es-en	de-en	es-en	fr-en	ru-en	en-de	en-es	en-ru
All Crowd-sourced Assessments	2,700	7,100	60,400	4,400	2,700	10,300	10,000	8,400
Post Quality Control	1,800	2,300	33,200	1,700	1,600	5,800	7,600	5,400
Distinct Translations	70	70	500	70	70	140	140	140

Table 1: Initial human assessment of nine language pairs, sample taken from all WMT-13 translation task participating systems.

Finally, the comparison made between the correlation of HTER (with human assessment) and those of other possible formulations, such as HBLEU, was unfortunately biased in favour of HTER, as Snover et al. (2006) report "Annotators were coached on how to minimize the edit rate. The coaching given to the annotators in order to minimize HTER, consisted mostly of teaching them which edits were considered by TER. The annotators also consulted with each other during training to compare various techniques they found to minimize HTER scores" (p. 227). A clear bias was therefore introduced into the comparison of correlations achieved by HTER and other human-targeted metrics, such as HBLEU for example, by coaching post-editors specifically to reduce HTER scores. Although bias in favour of HTER was identified in Snover et al. (2006), reporting "It is possible that performance of HTER is biased, as the targeted references were designed to minimize TER scores rather than BLEU scores" (p. 228), this was not highlighted or reconsidered when reporting correlations and drawing conclusions about the superiority of HTER over other metrics.

In this paper, we repeat the original experiments to re-assess HTER as a human assessment substitute. In doing so, we vastly expand experimental settings relative to the limitations of the original, as follows:

- Data sets include translations sampled from the output of 131 MT systems, operating at all levels of performance;
- Correlations of HTER with human assessment are reported for nine language pairs;
- Evaluation is based on between 15 and 80 human assessments per translation – to overcome human annotator agreement issues;
- Human-targeted reference translations are created without any coaching.

3 Re-evaluating Human-Targeted Metrics

As in evaluation of fully automatic metrics, evaluation of human-targeted metrics is by correlation with human assessment. Like the original evaluation of HTER, our re-evaluation also operates on the segment-level, with the performance of different possible human-targeted metric formulations measured by correlation of metric scores with segment-level human assessment. To compute correlations for human-targeted metrics, two distinct human annotations of each translation are required.

Firstly, we require a human assessment rating of the quality of each translation in order to compute correlations of each human-targeted metric. Our evaluation additionally requires manually created post-editions of each original MT output translation. This human post-edit is employed as the human-targeted reference translation for computing human-targeted metric scores for translations. We firstly describe the methodology employed for human assessment, before providing post-editing details.

3.1 Human Assessment of Translation Adequacy

Overall in this work, human assessment was carried out for two distinct data sets, firstly a nine language pair data set to re-evaluate HTER and subsequently an English to Spanish data set (see Section 4.2). Below we describe human assessment of the nine language pair data set.

Human direct assessment (DA) of the adequacy of translations was collected by means of a 100-point continuous rating scale via crowd-sourcing. Large numbers of repeat human judgments for translations were collected on Amazon Mechanical Turk, by way of re-implementation and minor adaptation of

Graham et al. (2015).[2] Translations were sampled at random from all systems competing in WMT-13 translation task for Czech-to-English (CS-EN), German-to-English (DE-EN), Spanish-to-English (ES-EN), French-to-English (FR-EN), Russian-to-English (RU-EN), English-to-German (EN-DE), English-to-Spanish (EN-ES), English-to-French (EN-FR) and English-to-Russian (EN-RU). Human assessors rated the adequacy of translations in a monolingual human evaluation by comparing the meaning of the original generic human reference translation (rendered in gray) with a given MT output translation (rendered in black) by stating the degree to which they agree that:[3]

The text in black adequately expresses the meaning of the text in gray in English.

Quality control was applied by comparison of ratings provided by each Mechanical Turk worker for pairs of original and degraded translations hidden within HITs, where each pair contained an original MT output translation along with a degraded version of that translation. Ratings of workers who were not reliable (or at any point during the course of data collection became unreliable) in their ability to accurately distinguish between the adequacy of degraded and original system output translations were omitted from the final data set.

Between 15 and 80 (22.4 on average) human assessments of adequacy were collected per translation and combined into a mean score for a given translation, after standardization of scores based on the individual annotators mean and standard deviation rating. Table 1 shows numbers of human assessments collected contributing to mean adequacy scores for translations.

3.2 Human-Targeted Reference Translations

A human-targeted reference translation was created for every translation in the data set by a known-reliable volunteer post-editor native in the target language of the MT output in question. All human post-editors were unaware of the purpose of the post-editing work and experiments. Annotators were shown the reference and MT output with post-editing instructions as follows:[4]

**Making as few changes as possible **, correct the hypothesis segment to make it

(a) have the same meaning as the reference segment;

(b) grammatically correct.

3.3 HTER Re-evaluation Results

Table 2 shows correlations achieved by a range of human-targeted metric formulations with human assessment, including human-targeted versions of segment-level BLEU (Papineni et al., 2002; Koehn et al., 2007), TER (Snover et al., 2006), WER, PER and CDER (Leusch and Ney, 2008). Correlations with human assessment achieved by HTER are, for three of the nine language pairs we evaluate, below a Pearson correlation of 0.6. In addition, comparison of correlations achieved by HTER and HBLEU reveal a *higher* correlation achieved by HBLEU for five of the nine language pairs we evaluate.

As recommended by Graham et al. (2015), we test for significance of difference in dependent correlations using Williams test, as shown in Figures 1 and 2. For two language pairs, DE-EN and FR-EN, correlations achieved by HBLEU with human assessment are significantly stronger than those achieved by HTER, and for all other language pairs, the difference in correlation achieved by HBLEU and HTER is not statistically significant.

Table 2 also includes correlations achieved by all metrics with human assessment when the original generic reference is employed. As expected correlations of generic-reference metrics are in all cases substantially lower than that of their human-targeted counterparts.

3.4 Possibility of Reference Bias in Monolingual MT Assessment

As mentioned in Section 1, in automatic evaluation of MT reference bias is a known problem, as high quality MT output can be unfairly penalized simply due to a lack of superficial similarity with the generic

[2]Complete implementation is made available at `https://github.com/ygraham/direct-assessment`

[3]All instructions were translated into the appropriate target language by a known-reliable native speaker.

[4]Due to space limitations, complete annotation instructions are provided at `https://github.com/ygraham/direct-assessment/post-editing-guidelines`

	CS-EN	DE-EN	ES-EN	FR-EN	RU-EN	EN-DE	EN-ES	EN-FR	EN-RU
H-TER	0.607	0.779	0.669	0.674	0.589	0.543	0.732	0.654	0.550
H-BLEU	0.612	**0.845**	0.664	**0.740**	0.588	**0.582**	0.710	0.637	0.573
H-CDER	0.603	0.828	**0.677**	0.635	**0.632**	0.579	0.718	**0.655**	**0.579**
H-WER	0.560	0.796	0.660	0.668	0.593	0.556	**0.733**	0.609	0.557
H-PER	**0.670**	0.757	0.650	0.651	0.552	0.467	0.679	0.619	0.543
TER	0.230	0.480	0.389	0.508	0.041	0.247	0.315	0.271	0.421
BLEU	0.153	0.429	**0.433**	0.389	0.040	**0.475**	**0.411**	**0.372**	**0.547**
CDER	**0.247**	**0.530**	0.426	0.409	**0.187**	0.353	0.363	0.238	0.401
PER	0.192	0.479	0.351	**0.553**	0.013	0.174	0.271	0.213	0.419
WER	0.198	0.489	0.382	0.425	0.065	0.273	0.325	0.216	0.384

Table 2: Correlation of segment-level human-targeted metric scores with human assessment and correlations of raw metrics with human assessment for a random sample of translations from WMT-13 translation task system submissions

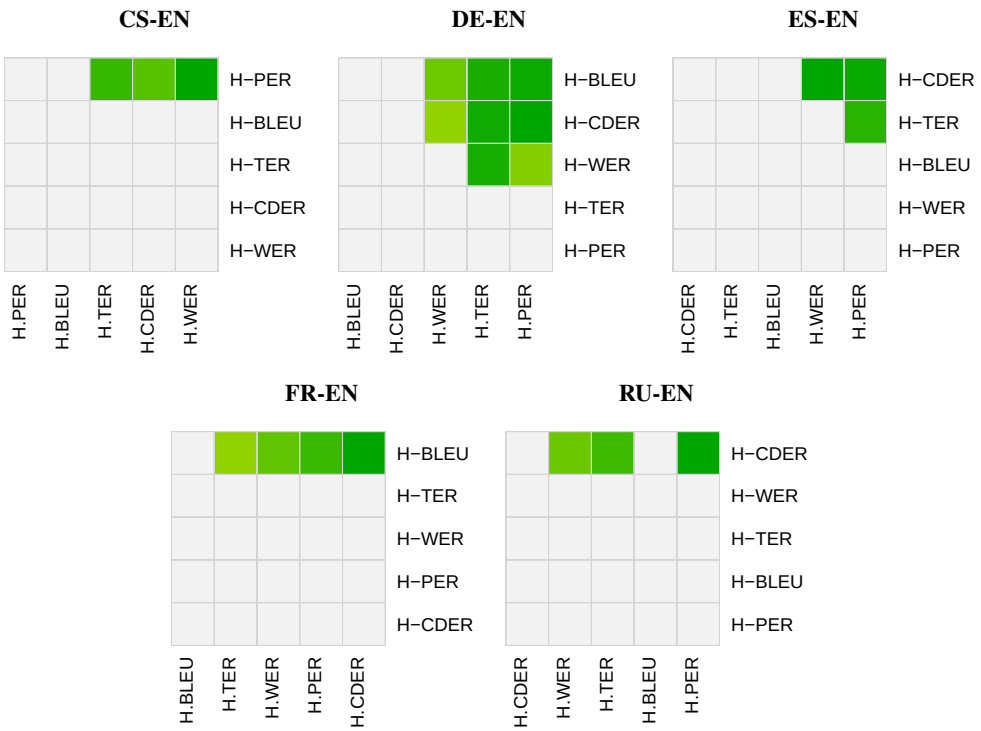

Figure 1: Significance test results (Williams test) of statistical significance of a difference in dependent correlations with human assessment for competing to-English human-targeted metrics; a green cell denotes a significant increase in correlation with human assessment for the metric in a given row over the metric in a given column at $p < 0.05$.

reference translation. Although certainly not to the same extreme, human assessment of MT that employs a reference translation could incur similar reference bias. Graham et al. (2013) provide discussion on how reference bias could exist for direct assessment of translation adequacy, and, as a solution, propose inclusion of a separate reference-free fluency assessment, to provide a component of the evaluation that cannot be biased in any way by a reference translation.

However, it is inevitably the case that resources available to conduct human evaluation are limited, and therefore a trade-off exists between adding fluency assessments of translations and numbers of translations we can feasibly include. For example, in our earlier human evaluation of HTER, it would have been possible to allocate resources to assessment of the fluency of translations, in addition to adequacy, but this would have come at the cost of reducing the size of the test set by approximately one half. Therefore,

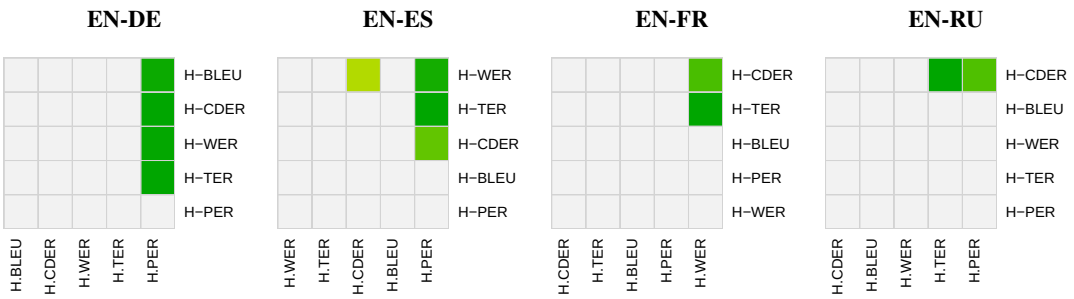

Figure 2: Significance test results (Williams test) of statistical significance of a difference in dependent correlations with human assessment for competing out-of-English human-targeted metrics; a green cell denotes a significant increase in correlation with human assessment for the metric in a given row over the metric in a given column at $p < 0.05$.

the decision to include fluency should be weighed against the degree to which reference bias is actually present in the adequacy evaluation. If there is strong reference bias, for example, it would be highly advisable to include both fluency and adequacy but for a smaller number of translations.

Fomicheva and Specia (2016) investigate bias in monolingual evaluation of MT and conclude reference bias to be a serious issue, with human annotators strongly biased by the reference translation provided. Their conclusion was based on estimation of confidence intervals for Kappa coefficients by means of an unconventional resampling technique, where samples used to estimate confidence intervals were smaller than the original sample size, drawn without replacement, and averaged; this diverges from standard methods of confidence interval estimation, such as bootstrap resampling. As a result, the reported confidence limits are unreliable, bringing their conclusions into question. We therefore consider it necessary to investigate the effect of reference bias with respect to the human evaluation we employ, to assess the likelihood of our own results with respect to the evaluation of HTER being contaminated by strong reference bias.[5]

Reference bias is the attribution of unfairly low scores to translations simply because they are not superficially similar to generic reference translations. In our evaluation of HTER, we employ DA adequacy scores by human assessment, where the MT output is compared to a generic reference translation. We will refer to DA scores collected in this way as $\text{DA}_{\text{gen-ref}}$. It is these gold standard $\text{DA}_{\text{gen-ref}}$ scores that potentially run the risk of introducing a strong reference bias into our evaluation of HTER.

At first, it might seem reasonable to attempt to measure this reference bias by comparison of $\text{DA}_{\text{gen-ref}}$ scores for translations with scores from an equivalent bilingual human evaluation, where the MT output is no longer compared to a reference translation, but is instead compared to the source language input and evaluated by a bilingual speaker. We refer to DA scores collected in this way as DA_{src}. However, in this case, DA_{src} scores could in fact include a different bias, one introduced by the fact that human assessors are now non-native in at least one of the required languages. To complicate things further, since non-native language skill levels will vary considerably from one human assessor to another, the degree of bias is likely to change considerably depending on particular language ability levels. Therefore, in addition to the potential of monolingual reference bias, we point out the real possibility in the case of DA_{src} of bias caused by reliance on bilingual human assessors not native in either the source or the target language. A comparison between DA_{src} and $\text{DA}_{\text{gen-ref}}$ scores for translations, therefore, would unfortunately not provide a reliable measurement of the monolingual reference bias, since such a comparison could be confounded by this additional bias.

Subsequently, we carry out a distinct comparison and motivate it as follows. Reference bias in $\text{DA}_{\text{gen-ref}}$ scores should cause unfairly low scores for only *some* translations, those that do not closely match generic reference translations but are in fact high quality translations. The difficulty in measur-

[5]This additional investigation was requested during peer review. Due to time limitations we only provide a preliminary investigation including a single language pair in this current publication. We hope to provide a more detailed study in the near future.

ing reference bias lies in separating the fair attribution of low scores to translations that do not closely match the reference and are *genuinely low quality* from those that also do not correspond closely to a given reference but nonetheless are in fact *high quality*. A comparison that would provide insight into the degree to which reference bias exists for $DA_{\text{gen-ref}}$ is one that compares $DA_{\text{gen-ref}}$ scores with those collected in a corresponding monolingual DA evaluation in which the surface similarity between MT output and reference translations is maximized for all translations in the test set. In this way, translations that have been unfairly penalized by $DA_{\text{gen-ref}}$ can be identified as having low $DA_{\text{gen-ref}}$ scores, when unfairly evaluated by comparison with a dissimilar generic reference, while at the same time achieving a high score when evaluated against a maximally similar reference translation. We therefore run DA with reference translations that have maximal surface similarity to MT output translations, by creating a human-targeted reference translation for each MT output.

3.4.1 DA with Source-Generated Human-Targeted References

In our earlier evaluation of HTER, human-targeted references were created by comparison with the generic reference. To guard against reference-bias being introduced in this current experiment, human-targeted references are now created without the use of the generic reference, instead by a bilingual examination of the original source language input and MT output only. We therefore employ a known-reliable bilingual native speaker of the source language with high fluency in the target language, as well as a known-reliable native speaker of the target language. In a first step, the bilingual post-edits the MT output, and in a second step, the post-edited MT output is checked for fluency in the target language by the monolingual in consultation with the bilingual to ensure human-targeted references remained faithful to the meaning of the source input.[6] We will refer to this additional set of DA scores as $DA_{\text{src-ht-ref}}$. In this way, translations that have been unfairly penalized by $DA_{\text{gen-ref}}$ can be identified as having low $DA_{\text{gen-ref}}$ scores, when unfairly evaluated by comparison with a dissimilar generic reference, while at the same time achieving a high $DA_{\text{src-ht-ref}}$ score when evaluated against a maximally similar reference translation.

3.4.2 Reference Bias in Monolingual MT Evaluation Experiment Results

Figure 3 (b) shows a scatter plot of $DA_{\text{src-ht-ref}}$ and $DA_{\text{gen-ref}}$ scores for our previous sample of RU-EN translations originally sampled from WMT-13. If a set of translations scored by $DA_{\text{gen-ref}}$ include reference bias, we would expect them to appear in the lower right quadrant of Figure 3 (b), as they will receive a low $DA_{\text{gen-ref}}$ score in combination with a high $DA_{\text{src-ht-ref}}$ score. As can be seen from the lower right portion of Figure 3 (b), only a small number of translations fall into the lower right quadrant, and the small number that do, all lie in very close proximity to neighboring upper-right and lower-left quadrants. Conversely, Figure 3 (a) shows the correspondence between $DA_{\text{src-ht-ref}}$ scores and the top-performing metric for that language pair, CDER,[7] where, in contrast, many translations, located in the lower right quadrant, receive an unfairly low CDER score while being scored highly by $DA_{\text{src-ht-ref}}$.

The lack of translations simultaneously receiving a low $DA_{\text{gen-ref}}$ score and high $DA_{\text{src-ht-ref}}$ score indicates that even though $DA_{\text{gen-ref}}$ employs a generic reference translation, scores are not strongly biased by the presence of this reference translation. This reinforces the validity of our human evaluation methodology, and should provide reassurance that in our direct assessment set-up, human assessors do in fact apply their human intelligence to the task, by reading a given translation and comparing its meaning to that of the generic reference, as instructed, as opposed to attributing scores to translations based on mere surface similarities between the MT output and generic reference translation.

4 Machine Translation Quality Estimation and Human-targeted Metrics

The now apparent lack of reliability of HTER as a valid substitute for human assessment raises doubts as to whether or not HTER scores should be relied upon as a gold standard representation for tasks such as QE, where the ultimate goal of systems is to predict translation quality. For example, previous evaluations of quality estimation systems have made the assumption that HTER provides a reliable human evaluation

[6] Post-editing guidelines are provided at `https://github.com/ygraham/direct-assessment/bilingual-postedit-guidelines`.

[7] Metric scores are standardized by the mean and standard deviation for ease of comparison.

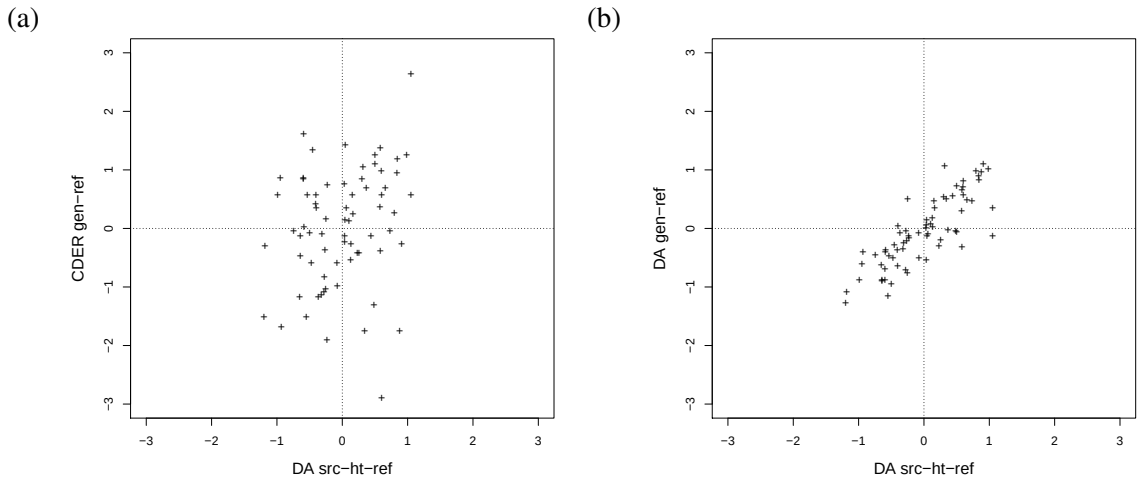

Figure 3: (a) Correspondence between DA scores collected by comparison with a source-generated human-targeted reference translation, DA src-ht-ref, where all translations benefit equally from close similarity to human-targeted references and automatic metric scores, CDER gen-ref, known to be biased by the generic reference and (b) the same instead compared to DA scores collected with comparison to generic reference translations, DA gen-ref. Lack of translations appearing in the lower right quadrant of (b) suggest DA human evaluation that employs a generic reference does not suffer from the strong reference bias known to be present in automatic metric scores (WMT-13 Russian to English).

substitute, but since we now know that HTER scores can be unrepresentative of translation quality, we rerun the previous evaluation provided by Bojar et al. (2013) and subsequently Graham (2015), instead employing DA human assessment as gold standard labels.

4.1 Human Annotation

Human-targeted reference translations were readily available in the WMT-13 shared task data set, while the human assessment method described in Section 3.1, was again applied to translations via crowdsourcing on Mechanical Turk, again applying strict quality control and computing mean scores for individual translations from a minimum of 15 human assessments. All human assessors who passed quality control by having significantly different scores for degraded versus genuine MT system output also showed no significant difference between mean rating scores for exact repeat translations.

Figure 4 (a) plots correspondence between HTER scores and human assessment, where the correlation of HTER with human assessment achieved is 0.678, almost at the same level as the strongest correlation in our nine-language pair evaluation. This data set therefore provides a best-case scenario for comparing differences in QE system rankings when HTER, as opposed to when human assessment, is employed as gold labels.

Before we compare QE system rankings, however, we carry out an analysis of a somewhat surprising relationship between HTER scores and human assessment for some translations shown in the scatter-plot in Figure 4 (a), where a number of translations appear to achieve a perfect HTER score of zero while at the same time cover a wide range of different human adequacy levels. Achieving a perfect HTER score means that the MT output was considered completely correct and required no post-editing whatsoever, so observing perfect HTER scores for translations that also receive low human assessment scores is a likely indication that something is amiss with either the post-editing carried out to create human-targeted references used to compute HTER scores or the human assessment.

Details of two example translations are provided in Figure 5. In Example 1, the source of error is the human post-editor (WMT-13 data set), since the MT output was in fact ungrammatical and despite this went uncorrected by the human post-editor, and this subsequently yielded an incorrect HTER score. Conversely, the error in Example 2 is caused by there being no direct translation of *straightforward* in Spanish and something unexpected subsequently taking place. Both the generic reference and the

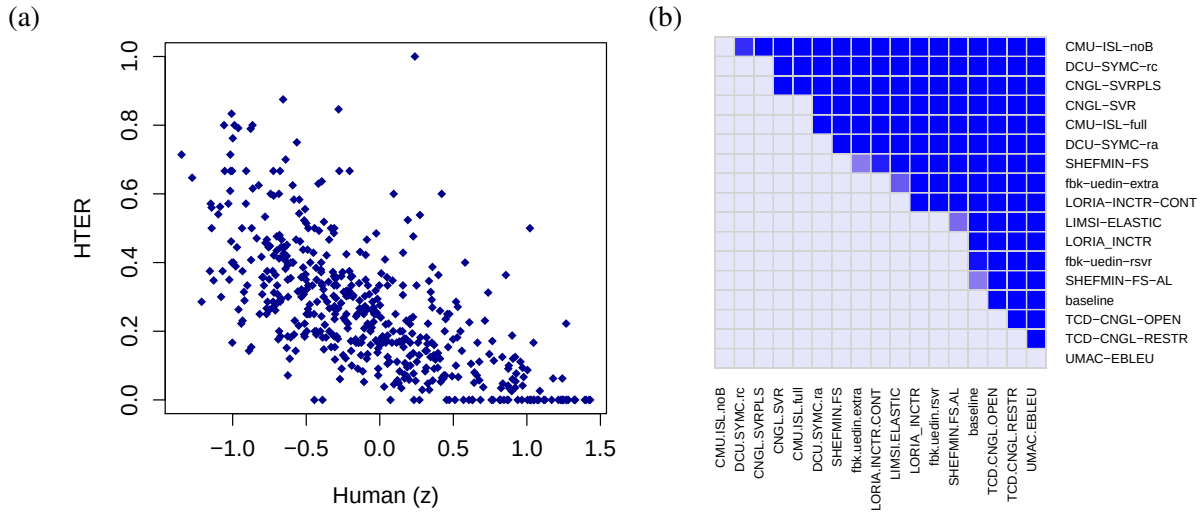

Figure 4: (a) Correspondence of HTER scores for translations in WMT-13 quality estimation Task 1.1 and human adequacy scores, and (b) significance test results (Williams test) of statistical significance of a difference in dependent correlations with human assessment for competing WMT-13 Task 1.1 quality estimation systems, a dark blue cell denotes a significant increase in correlation with human assessment for the QE system in a given row over the system in a given column at $p < 0.05$.

			HTER	Hum. z	Hum. raw
1.	source	Maybe we're more "Mexican" than I thought.			
	MT == HT ref.	*Tal vez estamos más "de México" de lo que se pensaba. *Perhaps we are more "from Mexico" than previously thought.*	0	−0.44	43%
	generic ref.	Quizás seamos más "mexicanos" de lo que pensaba. *Perhaps we are more "Mexican" than I thought.*			
2.	source	A straightforward man			
	MT == HT ref.	Un hombre sencillo *A simple man*	0	−0.39	43%
	generic ref.	Un hombre sincero *A sincere man*			

Figure 5: Example translations with a perfect HTER score (according to WMT-13 QE data set) and range of human assessment scores, where *denotes text ungrammatical in Spanish; generic ref. = the generic reference displayed to human assessors; MT = the MT output or human-targeted reference (HT ref.) employed by HTER (the latter being one and the same for translations receiving a perfect HTER score).

MT output could both be considered correct translations of the source while at the same time each of their meanings diverges somewhat from the other. Since the generic reference is employed for human assessment, the translation receives a low human adequacy score, despite it being a good translation of the original. Although human assessment is more valid than part-automatic metric scores, neither method is entirely impervious to other sources of error, therefore.

4.2 Quality Estimation Results

Table 3 shows correlations of QE system predictions with gold labels, where the rank order of systems diverges considerably when gold labels take the form of HTER scores as opposed to human assessment.

Significance test results for differences in correlation with human assessment for each pair of competing systems are provided in Figure 4 (b), where a new distinct outright winner of the task is identified that significantly outperforms all others, when evaluated with human assessment, as CMU-ISL-noB.

	Human Assessment		HTER			
QE System	r	Rank (r)	r	Rank (r)	MAE	Rank (MAE)
CMU-ISL-noB	0.571	1	0.516	5	0.138	7
DCU-SYMC-rc	0.557	2	0.595	1	0.135	5
CNGL-SVRPLS	0.553	3	0.560	4	0.133	3
CNGL-SVR	0.536	4	0.508	6	0.138	7
CMU-ISL-full	0.532	5	0.494	7	0.152	15
DCU-SYMC-ra	0.510	6	0.572	3	0.135	5
SHEFMIN-FS	0.489	7	0.575	2	0.124	1
fbk-uedin-extra	0.475	8	0.483	8	0.144	9
LORIA-INCTR-CONT	0.470	9	0.474	10	0.148	11
LIMSI-ELASTIC	0.459	10	0.475	9	0.133	3
LORIA_INCTR	0.452	11	0.461	13	0.148	11
fbk-uedin-rsvr	0.447	12	0.464	12	0.145	10
SHEFMIN-FS-AL	0.444	13	0.474	10	0.130	2
baseline	0.430	14	0.451	14	0.148	11
TCD-CNGL-OPEN	0.278	15	0.329	15	0.148	11
TCD-CNGL-RESTR	0.227	16	0.291	16	0.152	15
UMAC-EBLEU	0.017	17	0.113	17	0.170	17

Table 3: Pearson correlation of QE system predictions with HTER scores and correlation of system predictions with DA human adequacy scores for WMT-13 QE Task 1.1

The divergence in system rankings between evaluation of systems with HTER compared to DA indicates that even in the case that HTER correlates with human assessment to a relatively strong degree, 0.678, HTER is not a sufficient stand-in for human assessment. We subsequently recommend, where possible, employment of DA human assessment for evaluation of quality estimation systems, as it provides a valid human assessment without an automatic component, and has been shown to produce replicable sentence-level human assessment scores for translations (Graham et al., 2015).

5 Conclusion

Concerns were raised about the reliability of conclusions drawn in past experiments about HTER's superiority to other human-targeted metric formulations, and issues with respect to human post-editor coaching were highlighted as a source of bias, in addition to limited experiment settings. Subsequently, this motivated our re-evaluation of HTER extending to include a large number of systems across nine different language pairs. Results of our re-evaluation of HTER and several other possible metrics, including HBLEU, reveal significantly stronger correlations with human assessment for HBLEU compared to those achieved by HTER for two language pairs and no significant difference present in all other cases.

Further results showed correlations achieved by HTER may vary considerably across language pairs and in some cases are too low to provide a valid substitute for human assessment. As a test-case we replicated a previous quality estimation shared task and even when HTER scores correlate with human assessment at 0.678, they do not provide a valid human assessment substitute, with system rankings diverging considerably when DA human assessment is employed. We recommend DA human assessment for future evaluation of quality estimation.

Acknowledgments

We wish to thank Antonio Toral, Joachim Wagner and the anonymous reviewers for their feedback. This project has received funding from the European Union Horizon 2020 research and innovation programme under grant agreement 645452 (QT21) and 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.

References

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, August. 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, September. Association for Computational Linguistics.

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.

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

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 & Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

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*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1804–1813, Beijing, China. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, and Hieu Hoang. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Gregor Leusch and Hermann Ney. 2008. BLEUSP, INVWER, CDER: Three improved MT evaluation measures. In *Proceedings of NIST Metrics for Machine Translation 2008 Evaluation*, Honolulu, HI.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.

M. Snover, B. Dorr, R. Scwartz, J. Makhoul, and L. Micciula. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translaiton in the Americas*, pages 223–231, Boston, MA.