Journal of Artificial Intelligence Research 67 (2020) 653–672 Submitted 02/2019; published 03/2020

A Set of Recommendations for Assessing Human–Machine Parity in Language Translation

Samuel Läubli LAEUBLI@CL.UZH.CH
Institute of Computational Linguistics, University of Zurich

Sheila Castilho SHEILA.CASTILHO@ADAPTCENTRE.IE
ADAPT Centre, Dublin City University

Graham Neubig GNEUBIG@CS.CMU.EDU
Language Technologies Institute, Carnegie Mellon University

Rico Sennrich SENNRICH@CL.UZH.CH
Institute of Computational Linguistics, University of Zurich

Qinlan Shen QINLANS@CS.CMU.EDU
Language Technologies Institute, Carnegie Mellon University

Antonio Toral A.TORAL.RUIZ@RUG.NL
Center for Language and Cognition, University of Groningen

Abstract

The quality of machine translation has increased remarkably over the past years, to the degree that it was found to be indistinguishable from professional human translation in a number of empirical investigations. We reassess Hassan et al.'s 2018 investigation into Chinese to English news translation, showing that the finding of human–machine parity was owed to weaknesses in the evaluation design—which is currently considered best practice in the field. We show that the professional human translations contained significantly fewer errors, and that perceived quality in human evaluation depends on the choice of raters, the availability of linguistic context, and the creation of reference translations. Our results call for revisiting current best practices to assess strong machine translation systems in general and human–machine parity in particular, for which we offer a set of recommendations based on our empirical findings.

1. Introduction

Machine translation (MT) has made astounding progress in recent years thanks to improvements in neural modelling (Bahdanau, Cho, & Bengio, 2015; Sutskever, Vinyals, & Le, 2014; Vaswani et al., 2017), and the resulting increase in translation quality is creating new challenges for MT evaluation. Human evaluation remains the gold standard, but there are many design decisions that potentially affect the validity of such a human evaluation.

This paper is a response to two recent human evaluation studies in which some neural machine translation systems reportedly performed at (or above) the level of human translators

for news translation from Chinese to English (Hassan et al., 2018) and English to Czech (Bojar et al., 2018; Popel, 2018).

Both evaluations were based on current best practices in the field: they used a source-based direct assessment with non-expert annotators, using data sets and the evaluation protocol of the Conference on Machine Translation (WMT). While the results are intriguing, especially because they are based on best practices in MT evaluation, Bojar et al. (2018, p. 293) warn against taking their results as evidence for human–machine parity, and caution that "for well-resourced language pairs, an update of WMT evaluation style will be needed to keep up with the progress in machine translation." We concur that these findings have demonstrated the need to critically re-evaluate the design of human MT evaluation.

Our paper investigates three aspects of human MT evaluation, with a special focus on assessing human–machine parity: the choice of raters, the use of linguistic context, and the creation of reference translations. We focus on the data shared by Hassan et al. (2018), and empirically test to what extent changes in the evaluation design affect the outcome of the human evaluation.[1] We find that for all three aspects, human translations are judged more favourably, and significantly better than MT, when we make changes that we believe strengthen the evaluation design. Based on our empirical findings, we formulate a set of recommendations for human MT evaluation in general, and assessing human–machine parity in particular. All of our data are made publicly available for external validation and further analysis.[2]

2. Background

We first review current methods to assess the quality of machine translation system outputs, and highlight potential issues in using these methods to compare such outputs to translations produced by professional human translators.

2.1 Human Evaluation of Machine Translation

The evaluation of MT quality has been the subject of controversial discussions in research and the language services industry for decades due to its high economic importance. While automatic evaluation methods are particularly important in system development, there is consensus that a reliable evaluation should—despite high costs—be carried out by humans.

Various methods have been proposed for the human evaluation of MT quality (c.f. Castilho, Doherty, Gaspari, & Moorkens, 2018). What they have in common is that the MT output to be rated is paired with a translation hint: the source text or a reference translation. The MT output is then either adapted or scored with reference to the translation hint by human post-editors or raters, respectively.

1. Our results synthesise and extend those reported by Läubli, Sennrich, and Volk (2018) and Toral, Castilho, Hu, and Way (2018).
2. `https://github.com/ZurichNLP/mt-parity-assessment-data`

As part of the large-scale evaluation campaign at WMT, two primary evaluation methods have been used in recent years: relative ranking and direct assessment (Bojar, Federmann, et al., 2016). In the case of relative ranking, raters are presented with outputs from two or more systems, which they are asked to evaluate relative to each other (e.g., to determine system A is better than system B). Ties (e.g., system A is as good or as bad as system B) are typically allowed. Compared to absolute scores on Likert scales, data obtained through relative ranking show better inter- and intra-annotator agreement (Callison-Burch, Fordyce, Koehn, Monz, & Schroeder, 2007). However, they do not allow conclusions to be drawn about the order of magnitude of the differences, so that it is not possible to determine *how much* better system A was than system B.

This is one of the reasons why direct assessment has prevailed as an evaluation method more recently. In contrast to relative ranking, the raters are presented with one MT output at a time, to which they assign a score between 0 and 100. To increase homogeneity, each rater's ratings are standardised (Graham, Baldwin, Moffat, & Zobel, 2013). Reference translations serve as the basis in the context of WMT, and evaluations are carried out by monolingual raters. To avoid reference bias, the evaluation can be based on source texts instead, which presupposes bilingual raters, but leads to more reliable results overall (Bentivogli, Cettolo, Federico, & Federmann, 2018).

2.2 Assessing Human–Machine Parity

Hassan et al. (2018) base their claim of achieving human–machine parity on a source-based direct assessment as described in the previous section, where they found no significant difference in ratings between the output of their MT system and a professional human translation. Similarly, Bojar et al. (2018) report that the best-performing English to Czech system submitted to WMT 2018 (Popel, 2018) significantly outperforms the human reference translation. However, the authors caution against interpreting their results as evidence of human–machine parity, highlighting potential limitations of the evaluation.

In this study, we address three aspects that we consider to be particularly relevant for human evaluation of MT, with a special focus on testing human–machine parity: the choice of raters, the use of linguistic context, and the construction of reference translations.

Choice of Raters The human evaluation of MT output in research scenarios is typically conducted by crowd workers in order to minimise costs. Callison-Burch (2009) shows that aggregated assessments of bilingual crowd workers are "very similar" to those of MT developers, and Graham, Baldwin, Moffat, and Zobel (2017), based on experiments with data from WMT 2012, similarly conclude that with proper quality control, MT systems can be evaluated by crowd workers. Hassan et al. (2018) also use bilingual crowd workers, but the studies supporting the use of crowdsourcing for MT evaluation were performed with older MT systems, and their findings may not carry over to the evaluation of contemporary higher-quality neural machine translation (NMT) systems. In addition, the MT developers to which crowd workers were compared are usually not professional translators. We hypothesise that expert translators will provide more nuanced ratings than non-experts, and

that their ratings will show a higher difference between MT outputs and human translations.

Linguistic Context MT has been evaluated almost exclusively at the sentence level, owing to the fact that most MT systems do not yet take context across sentence boundaries into account. However, when machine translations are compared to those of professional translators, the omission of linguistic context—e. g., by random ordering of the sentences to be evaluated—does not do justice to humans who, in contrast to most MT systems, can and do take inter-sentential context into account (Voigt & Jurafsky, 2012; Wang, Tu, Way, & Liu, 2017). We hypothesise that an evaluation of sentences in isolation, as applied by Hassan et al. (2018), precludes raters from detecting translation errors that become apparent only when inter-sentential context is available, and that they will judge MT quality less favourably when evaluating full documents.

Reference Translations The human reference translations with which machine translations are compared within the scope of a human–machine parity assessment play an important role. Hassan et al. (2018) used all source texts of the WMT 2017 Chinese–English test set for their experiments, of which only half were originally written in Chinese; the other half were translated from English into Chinese. Since translated texts are usually simpler than their original counterparts (Laviosa-Braithwaite, 1998), they should be easier to translate for MT systems. Moreover, different human translations of the same source text sometimes show considerable differences in quality, and a comparison with an MT system only makes sense if the human reference translations are of high quality. Hassan et al. (2018), for example, had the WMT source texts re-translated as they were not convinced of the quality of the human translations in the test set. At WMT 2018, the organisers themselves noted that "the manual evaluation included several reports of ill-formed reference translations" (Bojar et al., 2018, p. 292). We hypothesise that the quality of the human translations has a significant effect on findings of human–machine parity, which would indicate that it is necessary to ensure that human translations used to assess parity claims need to be carefully vetted for their quality.

We empirically test and discuss the impact of these factors on human evaluation of MT in Sections 3–5. Based on our findings, we then distil a set of recommendations for human evaluation of strong MT systems, with a focus on assessing human–machine parity (Section 6).

2.3 Translations

We use English translations of the Chinese source texts in the WMT 2017 English–Chinese test set (Bojar et al., 2017) for all experiments presented in this article:

$\mathbf{H}_A$ The professional human translations in the dataset of Hassan et al. (2018).[1]

$\mathbf{H}_B$ Professional human translations that we ordered from a different translation vendor, which included a post-hoc native English check. We produced these only for the documents that were originally Chinese, as discussed in more detail in Section 5.2.

MT$_1$ The machine translations produced by Hassan et al.'s (2018) best system (Combo-6),[1] for which the authors found parity with H$_A$.

MT$_2$ The machine translations produced by Google's production system (Google Translate) in October 2017, as contained in Hassan et al.'s (2018) dataset.[1]

Statistical significance is denoted by * ($p \leq .05$), ** ($p \leq .01$), and *** ($p \leq .001$) throughout this article, unless otherwise stated.

3. Choice of Raters

Both professional and amateur evaluators can be involved in human evaluation of MT quality. However, from published work in the field (Doherty, 2017), it is fair to say that there is a tendency to "rely on students and amateur evaluators, sometimes with an undefined (or self-rated) proficiency in the languages involved, an unknown expertise with the text type" (Castilho et al., 2018, p. 23).

Previous work on evaluation of MT output by professional translators against crowd workers by Castilho et al. (2017) showed that for all language pairs (involving 11 languages) evaluated, crowd workers tend to be more accepting of the MT output by giving higher fluency and adequacy scores and performing very little post-editing. The authors argued that non-expert translators lack knowledge of translation and so might not notice subtle differences that make one translation more suitable than another, and therefore, when confronted with a translation that is hard to post-edit, tend to accept the MT rather than try to improve it.

3.1 Evaluation Protocol

We test for difference in ratings of MT outputs and human translations between experts and non-experts. We consider professional translators as experts, and both crowd workers and MT researchers as non-experts.[3]

We conduct a relative ranking experiment using one professional human (H$_A$) and two machine translations (MT$_1$ and MT$_2$), considering the native Chinese part of the WMT 2017 Chinese–English test set (see Section 5.2 for details). The 299 sentences used in the experiments stem from 41 documents, randomly selected from all the documents in the test set originally written in Chinese, and are shown in their original order. Raters are shown one sentence at a time, and see the original Chinese source alongside the three translations. The previous and next source sentences are also shown, in order to provide the annotator with local inter-sentential context.

Five raters—two experts and three non-experts—participated in the assessment. The experts were professional Chinese to English translators: one native in Chinese with a fluent

3. This terminology is not consistent with other literature, where MT researchers have been referred to as experts and crowd workers as non-experts (e. g., Callison-Burch, 2009).

Rank	Translators					
	All $n = 3873$		Experts $n = 1785$		Non-experts $n = 2088$	
1	H_A	1.939 *	H_A	2.247 *	H_A	1.324
2	MT_1	1.199 *	MT_1	1.197 *	MT_1	0.940 *
3	MT_2	−3.144	MT_2	−3.461	MT_2	−2.268

Table 1: Ranks and TrueSkill scores (the higher the better) of one human (H_A) and two machine translations (MT_1, MT_2) for evaluations carried out by expert and non-expert translators. An asterisk next to a translation indicates that this translation is significantly better than the one in the next rank at $p \leq .05$.

level of English, the other native in English with a fluent level of Chinese. The non-experts were NLP researchers native in Chinese, working in an English-speaking country.

The ratings are elicited with Appraise (Federmann, 2012). We derive an overall score for each translation (H_A, MT_1, and MT_2) based on the rankings. We use the TrueSkill method adapted to MT evaluation (Sakaguchi, Post, & Van Durme, 2014) following its usage at WMT15,[4] i. e., we run 1,000 iterations of the rankings recorded with Appraise followed by clustering (significance level $\alpha = 0.05$).

3.2 Results

Table 1 shows the TrueSkill scores for each translation resulting from the evaluations by expert and non-expert translators. We find that translation expertise affects the judgement of MT_1 and H_A, where the rating gap is wider for the expert raters.[5] This indicates that non-experts disregard translation nuances in the evaluation, which leads to a more tolerant judgement of MT systems and a lower inter-annotator agreement ($\kappa = 0.13$ for non-experts versus $\kappa = 0.254$ for experts).

It is worth noticing that, regardless of their expertise, the performance of human raters may vary over time. For example, performance may improve or decrease due to learning effects or fatigue, respectively (Gonzalez, Best, Healy, Kole, & Bourne, 2011). It is likely that such longitudinal effects are present in our data. They should be accounted for in future work, e. g., by using trial number as an additional predictor (Toral, Wieling, & Way, 2018).

4. `https://github.com/mjpost/wmt15`
5. As mentioned before, relative ranking mostly tells whether a translation is better than another but not by how much. The TrueSkill score is able to measure that difference, but may be difficult to interpret.

4. Linguistic Context

Another concern is the unit of evaluation. Historically, machine translation has primarily operated on the level of sentences, and so has machine translation evaluation. However, it has been remarked that human raters do not necessarily understand the intended meaning of a sentence shown out-of-context (Wu et al., 2016), which limits their ability to spot some mistranslations. Also, a sentence-level evaluation will be blind to errors related to textual cohesion and coherence.

While sentence-level evaluation may be good enough when evaluating MT systems of relatively low quality, we hypothesise that with additional context, raters will be able to make more nuanced quality assessments, and will also reward translations that show more textual cohesion and coherence. We believe that this aspect should be considered in evaluation, especially when making claims about human–machine parity, since human translators can and do take inter-sentential context into account (Voigt & Jurafsky, 2012; Wang et al., 2017).

4.1 Evaluation Protocol

We test if the availability of document-level context affects human–machine parity claims in terms of adequacy and fluency. In a pairwise ranking experiment, we show raters (i) isolated sentences and (ii) entire documents, asking them to choose the better (with ties allowed) from two translation outputs: one produced by a professional translator, the other by a machine translation system. We do not show reference translations as one of the two options is itself a human translation.

We use source sentences and documents from the WMT 2017 Chinese–English test set (see Section 2.3): documents are full news articles, and sentences are randomly drawn from these news articles, regardless of their position. We only consider articles from the test set that are native Chinese (see Section 5.2). In order to compare our results to those of Hassan et al. (2018), we use both their professional human (H_A) and machine translations (MT_1).

Each rater evaluates both sentences and documents, but never the same text in both conditions so as to avoid repetition priming (Francis & Sáenz, 2007). The order of experimental items as well as the placement of choices (H_A, MT_1; left, right) are randomised.

We use spam items for quality control (Kittur, Chi, & Suh, 2008): In a small fraction of items, we render one of the two options nonsensical by randomly shuffling the order of all translated words, except for 10 % at the beginning and end. If a rater marks a spam item as better than or equal to an actual translation, this is a strong indication that they did not read both options carefully.

We recruit professional translators (see Section 3) from `proz.com`, a well-known online market place for professional freelance translation, considering Chinese to English translators and native English revisers for the adequacy and fluency conditions, respectively. In each condition, four raters evaluate 50 documents (plus 5 spam items) and 104 sentences (plus

Context	N	Adequacy				Fluency			
		MT_1	Tie	H_A	p	MT_1	Tie	H_A	p
Sentence	208	49.5 %	9.1 %	41.4 %		31.7 %	17.3 %	51.0 %	**
Document	200	37.0 %	11.0 %	52.0 %	*	22.0 %	28.5 %	49.5 %	***

Table 2: Pairwise ranking results for machine (MT_1) against professional human translation (H_A) as obtained from blind evaluation by professional translators. Preference for MT_1 is lower when document-level context is available.

16 spam items). We use two non-overlapping sets of documents and two non-overlapping sets of sentences, and each is evaluated by two raters.

4.2 Results

Results are shown in Table 2. We note that sentence ratings from two raters are excluded from our analysis because of unintentional textual overlap with documents, meaning we cannot fully rule out that sentence-level decisions were informed by access to the full documents they originated from. Moreover, we exclude document ratings from one rater in the fluency condition because of poor performance on spam items, and recruit an additional rater to re-rate these documents.

We analyse our data using two-tailed Sign Tests, the null hypothesis being that raters do not prefer MT_1 over H_A or vice versa, implying human–machine parity. Following WMT evaluation campaigns that used pairwise ranking (e. g., Bojar et al., 2013), the number of successes x is the number of ratings in favour of H_A, and the number of trials n is the number of all ratings except for ties. Adding half of the ties to x and the total number of ties to n (Emerson & Simon, 1979) does not impact the significance levels reported in this section.

Adequacy raters show no statistically significant preference for MT_1 or H_A when evaluating isolated sentences ($x = 86, n = 189, p = .244$). This is in accordance with Hassan et al. (2018), who found the same in a source-based direct assessment experiment with crowd workers. With the availability of document-level context, however, preference for MT_1 drops from 49.5 to 37.0 % and is significantly lower than preference for human translation ($x = 104, n = 178, p < .05$). This evidences that document-level context cues allow raters to get a signal on adequacy.

Fluency raters prefer H_A over MT_1 both on the level of sentences ($x = 106, n = 172, p < .01$) and documents ($x = 99, n = 143, p < .001$). This is somewhat surprising given that increased fluency was found to be one of the main strengths of NMT (Bojar, Chatterjee, et al., 2016), as we further discuss in Section 5.1. The availability of document-level context decreases fluency raters' preference for MT_1, which falls from 31.7 to 22.0 %, without increasing their preference for H_A (Table 2).

Source	传统习俗引入新亮点“**2016孟兰文化节**”香港维园开幕敲锣打鼓的音乐、传统的小食、花俏的装饰、人群汹涌的现场。由香港潮属社团总会主办的“**2016孟兰文化节**”12日至14日在维多利亚公园举办，这是香港最盛大的一场孟兰胜会。
H_A	Traditional customs with new highlights - **2016 Ullam Cultural Festival** unveiled at Victoria Park in Hong Kong Music with drums and gongs, traditional snacks, fanciful decorations, and a chock-a-block crowd at the scene. The “**2016 Ullam Cultural Festival**” organized by the Federation of Hong Kong Chiu Chow Community Organizations will be held at Victoria Park from 12th to the 14th.
MT_1	Traditional customs introduce new bright spot “**2016 Ullambana Cultural Festival**” Hong Kong Victoria Park opening Gongs and drums music, traditional snacks, fancy decorations, the crowd surging scene. Organised by the Federation of Teochew Societies in Hong Kong, the “**2016 Python Cultural Festival**” is held at Victoria Park from 12 to 14 July.

Table 3: Two consecutive sentences of a Chinese news article as translated into English by a professional human translator (H_A) and a machine translation system (MT_1). Emphasis added.

4.3 Discussion

Our findings emphasise the importance of linguistic context in human evaluation of MT. In terms of adequacy, raters assessing documents as a whole show a significant preference for human translation, but when assessing single sentences in random order, they show no significant preference for human translation.

Document-level evaluation exposes errors to raters which are hard or impossible to spot in a sentence-level evaluation, such as coherent translation of named entities. The example in Table 3 shows the first two sentences of a Chinese news article as translated by a professional human translator (H_A) and Hassan et al.'s (2018) NMT system (MT_1). When looking at both sentences (document-level evaluation), it can be seen that MT_1 uses two different translations to refer to a cultural festival, “2016孟兰文化节”, whereas the human translation uses only one. When assessing the second sentence out of context (sentence-level evaluation), it is hard to penalise MT_1 for producing “2016 Python Cultural Festival”, particularly for fluency raters without access to the corresponding source text. For further examples, see Section 5.1 and Table 6.

5. Reference Translations

Yet another relevant element in human evaluation is the reference translation used. This is the focus of this section, where we cover two aspects of reference translations that can have an impact on evaluation: quality and directionality.

5.1 Quality

Because the translations are created by humans, a number of factors could lead to compromises in quality:

Errors in Understanding: If the translator is a non-native speaker of the source language, they may make mistakes in interpreting the original message. This is particularly true if the translator does not normally work in the domain of the text, e. g., when a translator who normally works on translating electronic product manuals is asked to translate news.

Errors in Fluency: If the translator is a non-native speaker of the target language, they might not be able to generate completely fluent text. This similarly applies to domain-specific terminology.

Limited Resources: Unlike computers, human translators have limits in time, attention, and motivation, and will generally do a better job when they have sufficient time to check their work, or are particularly motivated to do a good job, such as when doing a good job is necessary to maintain their reputation as a translator.

Effects of Post-editing: In recent years, a large number of human translation jobs are performed by post-editing MT output, which can result in MT artefacts remaining even after manual post-editing (Castilho, Resende, & Mitkov, 2019; Daems, Vandepitte, Hartsuiker, & Macken, 2017; Toral, 2019).

In this section, we examine the effect of the quality of underlying translations on the conclusions that can be drawn with regards to human–machine parity. We first do an analysis on (i) how the source of the human translation affects claims of human–machine parity, and (ii) whether significant differences exist between two varieties of human translation. We follow the same protocol as in Section 4.1, having 4 professional translators per condition, evaluate the translations for adequacy and fluency on both the sentence and document level.[6]

The results are shown in Table 4. From this, we can see that the human translation H_B, which was aggressively edited to ensure target fluency, resulted in lower adequacy (Table 4b). With more fluent and less accurate translations, raters do not prefer human over machine translation in terms of adequacy (Table 4a), but have a stronger preference for human translation in terms of fluency (compare Tables 4a and 2). In a direct comparison of the two human translations (Table 4b), we also find that H_A is considered significantly more adequate than H_B, while there is no significant difference in fluency.

To achieve a finer-grained understanding of what errors the evaluated translations exhibit, we perform a categorisation of 150 randomly sampled sentences based on the classification used by Hassan et al. (2018).[7] We expand the classification with a Context category, which we use to mark errors that are only apparent in larger context (e. g., regarding poor register choice, or coreference errors), and which do not clearly fit into one of the other categories.

6. Translators were recruited from `proz.com`.
7. Hassan et al.'s (2018) classification is in turn based on, but significantly different than that proposed by Vilar, Xu, Luis Fernando, and Ney (2006).

Context	N	Adequacy				Fluency			
		MT_1	Tie	H_B	p	MT_1	Tie	H_B	p
Sentence	416	34.6 %	18.8 %	46.6 %		20.7 %	21.2 %	58.2 %	***
Document	200	41.0 %	9.0 %	50.0 %		18.0 %	21.0 %	61.0 %	***

(a) Machine translation MT_1 against professional human translation H_B

Context	N	Adequacy				Fluency			
		H_A	Tie	H_B	p	H_A	Tie	H_B	p
Sentence	416	56.7 %	10.6 %	32.7 %	***	40.4 %	24.0 %	35.6 %	
Document	200	64.0 %	9.0 %	27.0 %	***	34.0 %	22.0 %	44.0 %	

(b) Professional human translation H_A against professional human translation H_B

Table 4: Pairwise ranking results for one machine (MT_1) and two professional human translations (H_A, H_B) as obtained from blind evaluation by professional translators.

Hassan et al. (2018) perform this classification only for the machine-translated outputs, and thus the natural question of whether the mistakes that humans and computers make are qualitatively different is left unanswered. Our analysis was performed by one of the co-authors who is a bi-lingual native Chinese/English speaker. Sentences were shown in the context of the document, to make it easier to determine whether the translations were correct based on the context. The analysis was performed on one machine translation (MT_1) and two human translation outputs (H_A, H_B), using the same 150 sentences, but blinding their origin by randomising the order in which the documents were presented. We show the results of this analysis in Table 5.

From these results, we can glean a few interesting insights. First, we find significantly larger numbers of errors of the categories of Incorrect Word and Named Entity in MT_1, indicating that the MT system is less effective at choosing correct translations for individual words than the human translators. An example of this can be found in Table 6a, where we see that the MT system refers to a singular "point of view" and translates "线路" (channel, route, path) into the semantically similar but inadequate "lines". Interestingly, MT_1 has significantly more Word Order errors, one example of this being shown in Table 6b, with the relative placements of "at the end of last year" (去年年底) and "stop production" (停产). This result is particularly notable given previous reports that NMT systems have led to great increases in reordering accuracy compared to previous statistical MT systems (Bentivogli, Bisazza, Cettolo, & Federico, 2016; Neubig, Morishita, & Nakamura, 2015), demonstrating that the problem of generating correctly ordered output is far from solved even in very strong NMT systems. Moreover, H_B had significantly more Missing Word (Semantics) errors than both H_A ($p < .001$) and MT_1 ($p < .001$), an indication that the proofreading

Error Category	**Errors**			**Significance**		
	H_A	H_B	MT_1	$H_A - H_B$	$H_A - MT_1$	$H_B - MT_1$
Incorrect Word	51	52	85		***	***
Semantics	33	36	48			
Grammaticality	18	16	37		**	**
Missing Word	37	69	56	***	*	
Semantics	22	62	34	***		***
Grammaticality	15	7	22			**
Named Entity	16	19	30		*	
Person	1	10	10	*	*	
Location	5	4	6			
Organization	4	4	8			
Event	1	1	3			
Other	5	1	7			
Word Order	1	4	17		***	**
Factoid	1	1	6			
Word Repetition	2	4	4			
Collocation	15	18	27			
Unknown Words/Misspellings	0	1	0			
Context (Register, Coreference, etc.)	6	9	12			
Any	81	103	118	*	***	
Total	129	177	237	**	***	**

Table 5: Classification of errors in machine translation MT_1 and two professional human translation outputs H_A and H_B. Errors represent the number of sentences (out of $N = 150$) that contain at least one error of the respective type. We also report the number of sentences that contain at least one error of any category (Any), and the total number of error categories present in all sentences (Total). Statistical significance is assessed with Fisher's exact test (two-tailed) for each pair of translation outputs.

Source	在目前较为主流的**观点**中，番薯的引进主要有三条**线路**。
H_A	Currently more mainstream **perspectives** point to three **channels** for the introduction of sweet potatoes.
H_B	There are currently three **theories** in regards to how the sweet potato was introduced.
MT_1	In the current more mainstream **point of view**, the introduction of sweet potato has three main **lines**.

(a) Incorrect Word

Source	该企业位于青岛老城区的厂区**去年年底**全面**停产**,环保搬迁至平度的新厂区。
H_A	This corporation, situated in the factory area of Old Town of Qingdao, **stopped its production** lines **at the end of last year**.
H_B	The same company had a plant in Qingdao's old town but was **shut down last year**.
MT_1	The enterprise is located in the old city of Qingdao plant **at the end of last year** to **stop production**.

(b) Reordering

Source	据知情人士透露，**近期**，苏宁高层与苹果公司频频见面，目的就是为了准备充足的货源。
H_A	Insider disclosure revealed that the upper management of Suning and Apple has met frequently **in recent times** for the purpose of preparing enough resources.
H_B	According to informed sources, Suning executives met frequently with Apple in order to prepare enough stock.
MT_1	According to people familiar with the matter, **recently**, Suning executives have been meeting with Apple frequently in order to prepare an adequate supply.

(c) Missing Word (Semantics)

Table 6: Qualitative examples of differences in types of errors found in machine (MT_1) and two professional human translations (H_A, H_B). Emphasis added.

Source	张彬彬和**家人**聚少离多... **父母**说... **张彬彬**很少说**自己的辛苦**，更多的是跟**父母**聊些开心的事。
H_A	Zhang Binbin spends little time with **family**... **Her parents** said... Zhang Binbin seldom said **she found things difficult**. More often, **she** would chat about happy things with parents.
H_B	Zhang Binbin saw **her family** less... **Her parents** said... **she** would seldom talk about **her hardship** and would mostly talk about something happy with **her** parents
MT_1	Zhang Binbin and **his family** gathered less... Parents said... Zhang Binbin rarely said **their hard work**, more with **their parents** to talk about something happy.

(d) Context

Table 6: (Continued from previous page.)

process resulted in drops of content in favour of fluency. An example of this is shown in Table 6c, where H_B dropped the information that the meetings between Suning and Apple were *recently* (近期) held. Finally, while there was not a significant difference, likely due to the small number of examples overall, it is noticeable that MT_1 had a higher percentage of Collocation and Context errors, which indicate that the system has more trouble translating words that are dependent on longer-range context. Similarly, some Named Entity errors are also attributable to translation inconsistencies due to lack of longer-range context. Table 6d shows an example where we see that the MT system was unable to maintain a consistently gendered or correct pronoun for the female Olympic shooter Zhang Binbin (张彬彬).

Apart from showing qualitative differences between the three translations, the analysis also supports the finding of the pairwise ranking study: H_A is both preferred over MT_1 in the pairwise ranking study, and exhibits fewer translation errors in our error classification. H_B has a substantially higher number of missing words than the other two translations, which agrees with the lower perceived adequacy in the pairwise ranking.

However, the analysis not only supports the findings of the pairwise ranking study, but also adds nuance to it. Even though H_B has the highest number of deletions, and does worse than the other two translations in a pairwise adequacy ranking, it is similar to H_A, and better than MT_1, in terms of most other error categories.

5.2 Directionality

Translation quality is also affected by the nature of the source text. In this respect, we note that from the 2,001 sentences in the WMT 2017 Chinese–English test set, half were originally written in Chinese; the remaining half were originally written in English and then manually translated into Chinese. This Chinese reference file (half original, half translated) was then manually translated into English by Hassan et al. (2018) to make up the reference

Rank	Original Language					
	Both $n = 6675$		Chinese $n = 3873$		English $n = 2802$	
1	H_A	1.587 *	H_A	1.939 *	MT_1	1.059
2	MT_1	1.231 *	MT_1	1.199 *	H_A	0.772 *
3	MT_2	-2.819	MT_2	-3.144	MT_2	-1.832

Table 7: Ranks of the translations given the original language of the source side of the test set shown with their TrueSkill score (the higher the better). An asterisk next to a translation indicates that this translation is significantly better than the one in the next rank at $p \leq .05$.

for assessing human–machine parity. Therefore, 50 % of the reference comprises direct English translations from the original Chinese, while 50 % are English translations from the human-translated file from English into Chinese, i. e., backtranslations of the original English.

According to Laviosa (1998), translated texts differ from their originals in that they are simpler, more explicit, and more normalised. For example, the synonyms used in an original text may be replaced by a single translation. These differences are referred to as translationese, and have been shown to affect translation quality in the field of machine translation (Castilho et al., 2019; Daems, De Clercq, & Macken, 2017; Kurokawa, Goutte, & Isabelle, 2009; Toral, 2019).

We test whether translationese has an effect on assessing parity between translations produced by humans and machines, using relative rankings of translations in the WMT 2017 Chinese–English test set by five raters (see Section 3). Our hypothesis is that the difference between human and machine translation quality is smaller when source texts are translated English (translationese) rather than original Chinese, because a translationese source text should be simpler and thus easier to translate for an MT system. We confirm Laviosa's observation that "translationese" Chinese (that started as English) exhibits less lexical variety than "natively" Chinese text and demonstrate that translationese source texts are generally easier for MT systems to score well on.

Table 7 shows the TrueSkill scores for translations (H_A, MT_1, and MT_2) of the entire test set (Both) versus only the sentences originally written in Chinese or English therein. The human translation H_A outperforms the machine translation MT_1 significantly when the original language is Chinese, while the difference between the two is not significant when the original language is English (i. e., translationese input).

We also compare the two subsets of the test set, original and translationese, using type-token ratio (TTR). Our hypothesis is that the TTR will be smaller for the translationese subset, thus its simpler nature getting reflected in a less varied use of language. While both subsets contain a similar number of sentences (1,001 and 1,000), the Chinese subset contains

more tokens (26,468) than its English counterpart (22,279). We thus take a subset of the Chinese (840 sentences) containing a similar amount of words to the English data (22,271 words). We then calculate the TTR for these two subsets using bootstrap resampling. The TTR for Chinese ($M = 0.1927$, $SD = 0.0026$, 95 % confidence interval $[0.1925, 0.1928]$) is 13 % higher than that for English ($M = 0.1710$, $SD = 0.0025$, 95 % confidence interval $[0.1708, 0.1711]$).

Our results show that using translationese (Chinese translated from English) rather than original source texts results in higher scores for MT systems in human evaluation, and that the lexical variety of translationese is smaller than that of original text.

6. Recommendations

Our experiments in Sections 3–5 show that machine translation quality has not yet reached the level of professional human translation, and that human evaluation methods which are currently considered best practice fail to reveal errors in the output of strong NMT systems. In this section, we recommend a set of evaluation design changes that we believe are needed for assessing human–machine parity, and will strengthen the human evaluation of MT in general.

(R1) Choose professional translators as raters. In our blind experiment (Section 3), non-experts assess parity between human and machine translation where professional translators do not, indicating that the former neglect more subtle differences between different translation outputs.

(R2) Evaluate documents, not sentences. When evaluating sentences in random order, professional translators judge machine translation more favourably as they cannot identify errors related to textual coherence and cohesion, such as different translations of the same product name. Our experiments show that using whole documents (i. e., full news articles) as unit of evaluation increases the rating gap between human and machine translation (Section 4).

(R3) Evaluate fluency in addition to adequacy. Raters who judge target language fluency without access to the source texts show a stronger preference for human translation than raters with access to the source texts (Sections 4 and 5.1). In all of our experiments, raters prefer human translation in terms of fluency while, just as in Hassan et al.'s (2018) evaluation, they find no significant difference between human and machine translation in sentence-level adequacy (Tables 2 and 4a). Our error analysis in Table 6 also indicates that MT still lags behind human translation in fluency, specifically in grammaticality.

(R4) Do not heavily edit reference translations for fluency. In professional translation workflows, texts are typically revised with a focus on target language fluency after an initial translation step. As shown in our experiment in Section 5.1, aggressive revision can make translations more fluent but less accurate, to the degree that they become indistinguishable from MT in terms of accuracy (Table 4a).

(R5) Use original source texts. Raters show a significant preference for human over machine translations of texts that were originally written in the source language, but not for source texts that are translations themselves (Section 5.2). Our results are further evidence that translated texts tend to be simpler than original texts, and in turn easier to translate with MT.

Our work empirically strengthens and extends the recommendations on human MT evaluation in previous work (Läubli et al., 2018; Toral, Castilho, et al., 2018), some of which have meanwhile been adopted by the large-scale evaluation campaign at WMT 2019 (Barrault et al., 2019): the new evaluation protocol uses original source texts only (R5) and gives raters access to document-level context (R2). The findings of WMT 2019 provide further evidence in support of our recommendations. In particular, human English to Czech translation was found to be significantly better than MT (Barrault et al., 2019, p. 28); the comparison includes the same MT system (`CUNI-Transformer-T2T-2018`) which outperformed human translation according to the previous protocol (Bojar et al., 2018, p. 291). Results also show a larger difference between human translation and MT in document-level evaluation.[8]

We note that in contrast to WMT, the judgements in our experiments are provided by a small number of human raters: five in the experiments of Sections 3 and 5.2, four per condition (adequacy and fluency) in Section 4, and one in the fine-grained error analysis presented in Section 5.1. Moreover, the results presented in this article are based on one text domain (news) and one language direction (Chinese to English), and while a large-scale evaluation with another language pair supports our findings (see above), further experiments with more languages, domains, and raters will be required to increase their external validity.

7. Conclusion

We compared professional human Chinese to English translations to the output of a strong MT system. In a human evaluation following best practices, Hassan et al. (2018) found no significant difference between the two, concluding that their NMT system had reached parity with professional human translation. Our blind qualitative analysis, however, showed that the machine translation output contained significantly more incorrect words, omissions, mistranslated names, and word order errors.

Our experiments show that recent findings of human–machine parity in language translation are owed to weaknesses in the design of human evaluation campaigns. We empirically tested alternatives to what is currently considered best practice in the field, and found that the choice of raters, the availability of linguistic context, and the creation of reference translations have a strong impact on perceived translation quality. As for the choice of raters, professional translators showed a significant preference for human translation, while non-expert raters did not. In terms of linguistic context, raters found human translation significantly more accurate than machine translation when evaluating full documents, but

8. Specifically, the absolute difference between HUMAN and `CUNI-Transformer-T2T-2018` in terms of average standardized human scores is 11–22% for segment-level evaluation, 24% for segment-level evaluation with document-level context, and 39% for document-level evaluation (Barrault et al., 2019, p. 28).

not when evaluating single sentences out of context. They also found human translation significantly more fluent than machine translation, both when evaluating full documents and single sentences. Moreover, we showed that aggressive editing of human reference translations for target language fluency can decrease adequacy to the point that they become indistinguishable from machine translation, and that raters found human translations significantly better than machine translations of original source texts, but not of source texts that were translations themselves.

Our results strongly suggest that in order to reveal errors in the output of strong MT systems, the design of MT quality assessments with human raters should be revisited. To that end, we have offered a set of recommendations, supported by empirical data, which we believe are needed for assessing human–machine parity, and will strengthen the human evaluation of MT in general. Our recommendations have the aim of increasing the validity of MT evaluation, but we are aware of the high cost of having MT evaluation done by professional translators, and on the level of full documents. We welcome future research into alternative evaluation protocols that can demonstrate their validity at a lower cost.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR.* San Diego, CA.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., ... Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of WMT* (pp. 1–61). Florence, Italy.

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of EMNLP* (pp. 257–267). Austin, Texas.

Bentivogli, L., Cettolo, M., Federico, M., & Federmann, C. (2018). Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment. In *Proceedings of IWSLT* (pp. 62–69). Bruges, Belgium.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., ... Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of WMT* (pp. 1–44). Sofia, Bulgaria.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., ... Turchi, M. (2017). Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of WMT* (pp. 169–214). Copenhagen, Denmark.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., ... Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of WMT* (pp. 131–198). Berlin, Germany.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., ... Monz, C. (2018). Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of WMT* (pp. 272–307). Belgium, Brussels.

Bojar, O., Federmann, C., Haddow, B., Koehn, P., Post, M., & Specia, L. (2016). Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In *Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem* (pp.

27–36). Portoroz, Slovenia.

Callison-Burch, C. (2009). Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of EMNLP* (pp. 286–295). Singapore.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-) evaluation of machine translation. In *Proceedings of WMT* (pp. 136–158). Prague, Czech Republic.

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In *Translation Quality Assessment: From Principles to Practice* (Vol. 1, pp. 9–38). Springer International Publishing.

Castilho, S., Moorkens, J., Gaspari, F., Way, A., Georgakopoulou, P., Gialama, M., ... Sennrich, R. (2017). Crowdsourcing for NMT evaluation: Professional translators versus the crowd. In *Proceedings of Translating and the Computer 39.* London, UK.

Castilho, S., Resende, N., & Mitkov, R. (2019). What Influences the Features of Post-editese? A Preliminary Study. In *Proceedings of HiT-IT* (pp. 19–27). Varna, Bulgaria.

Daems, J., De Clercq, O., & Macken, L. (2017). Translationese and Post-editese: How comparable is comparable quality? *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, *16*, 89–103.

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2017). Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta*, *62*(2), 245–270.

Doherty, S. (2017). Issues in human and automatic translation quality assessment. In *Human issues in translation technology* (pp. 131–148). Routledge.

Emerson, J. D., & Simon, G. A. (1979). Another Look at the Sign Test When Ties Are Present: The Problem of Confidence Intervals. *The American Statistician*, *33*(3), 140–142.

Federmann, C. (2012). Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, *98*, 25–35. (Code available at `https://github.com/cfedermann/Appraise`.)

Francis, W. S., & Sáenz, S. P. (2007). Repetition priming endurance in picture naming and translation: Contributions of component processes. *Memory & Cognition*, *35*(3), 481–493.

Gonzalez, C., Best, B., Healy, A. F., Kole, J. A., & Bourne, L. E. (2011). A cognitive modeling account of simultaneous learning and fatigue effects. *Cognitive Systems Research*, *12*(1), 19–32.

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013). Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse* (pp. 33–41). Sofia, Bulgaria.

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone? *Natural Language Engineering*, *23*(1), 3–30.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... Zhou, M. (2018). Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint 1803.05567*. (Data available at `http://aka.ms/Translator-HumanParityData`.)

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing User Studies with Mechanical

Turk. In *Proceedings of CHI* (pp. 453–456). Florence, Italy.

Kurokawa, D., Goutte, C., & Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII* (pp. 81–88).

Laviosa, S. (1998). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta*, *43*(4), 557–570.

Laviosa-Braithwaite, S. (1998). Universals of translation. In *Routledge Encyclopedia of Translation Studies* (pp. 288–291). Routledge.

Läubli, S., Sennrich, R., & Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of EMNLP* (pp. 4791–4796). Brussels, Belgium.

Neubig, G., Morishita, M., & Nakamura, S. (2015). Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015. In *Proceedings of WAT2015.* Kyoto, Japan.

Popel, M. (2018). CUNI Transformer Neural MT System for WMT18. In *Proceedings of WMT* (pp. 486–491). Brussels, Belgium.

Sakaguchi, K., Post, M., & Van Durme, B. (2014). Efficient Elicitation of Annotations for Human Evaluation of Machine Translation. In *Proceedings of WMT* (pp. 1–11). Baltimore, MD.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Proceedings of NIPS* (pp. 3104–3112). Montreal, Canada.

Toral, A. (2019). Post-editese: an Exacerbated Translationese. In *Proceedings of MT Summit* (pp. 273–281). Dublin, Ireland: European Association for Machine Translation.

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of WMT* (pp. 113–123). Brussels, Belgium.

Toral, A., Wieling, M., & Way, A. (2018). Post-editing Effort of a Novel With Statistical and Neural Machine Translation. *Frontiers in Digital Humanities*, *5*, 9.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In *Proceedings of NIPS* (pp. 5998–6008). Long Beach, CA.

Vilar, D., Xu, J., Luis Fernando, D., & Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of LREC* (pp. 697–702).

Voigt, R., & Jurafsky, D. (2012). Towards a Literary Machine Translation: The Role of Referential Cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature* (pp. 18–25). Montrèal, Canada.

Wang, L., Tu, Z., Way, A., & Liu, Q. (2017). Exploiting Cross-Sentence Context for Neural Machine Translation. In *Proceedings of EMNLP* (pp. 2826–2831). Copenhagen, Denmark.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint 1609.08144*.