Reassessing Claims of Human Parity and Super-Human Performance in Machine Translation at WMT 2019

Toral, Antonio

Published in:
Proceedings of the 22nd Annual Conference of the European Association for Machine Translation

Reassessing Claims of Human Parity and Super-Human Performance in Machine Translation at WMT 2019

Antonio Toral
Center for Language and Cognition
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

Abstract

We reassess the claims of human parity and super-human performance made at the news shared task of WMT 2019 for three translation directions: English→German, English→Russian and German→English. First we identify three potential issues in the human evaluation of that shared task: (i) the limited amount of intersentential context available, (ii) the limited translation proficiency of the evaluators and (iii) the use of a reference translation. We then conduct a modified evaluation taking these issues into account. Our results indicate that all the claims of human parity and super-human performance made at WMT 2019 should be refuted, except the claim of human parity for English→German. Based on our findings, we put forward a set of recommendations and open questions for future assessments of human parity in machine translation.

1 Introduction

The quality of the translations produced by machine translation (MT) systems has improved considerably since the adoption of architectures based on neural networks (Bentivogli et al., 2016). To the extent that, in the last two years, there have been claims of MT systems reaching human parity and even super-human performance (Hassan et al., 2018; Bojar et al., 2018; Barrault et al., 2019). Following Hassan et al. (2018), we consider that human parity is achieved for a given task *t* if the performance attained by a computer on *t* is equivalent to that of a human, i.e. there is no significant difference between the performance obtained by human and by machine. Super-human performance is achieved for *t* if the performance achieved by a computer is significantly better than that of a human.

Two claims of human parity in MT were reported in 2018. One by Microsoft, on news translation for Chinese→English (Hassan et al., 2018), and another at the news translation task of WMT for English→Czech (Bojar et al., 2018), in which MT systems Uedin (Haddow et al., 2018) and Cuni-Transformer (Kocmi et al., 2018) reached human parity and super-human performance, respectively. In 2019 there were additional claims at the news translation task of WMT (Barrault et al., 2019): human parity for German→English, by several of the submitted systems, and for English→Russian, by system Facebook-FAIR (Ng et al., 2019), as well as super-human performance for English→German, again by Facebook-FAIR.

The claims of human parity and super-human performance in MT made in 2018 (Hassan et al., 2018; Bojar et al., 2018) have been since refuted given three issues in their evaluation setups (Läubli et al., 2018; Toral et al., 2018): (i) part of the source text of the test set was not original text but translationese, (ii) the sentences were evaluated in isolation, and (iii) the evaluation was not conducted by translators. However, the evaluation setup of WMT 2019 was modified to address some of these issues: the first issue (translationese) was fully addressed, while the second (sentences evaluated in isolation) was partially addressed, as we will motivate in Section 2.1, whereas the third (human evaluation conducted by non-translators) was not acted upon. Given that some of the issues that led to refute the claims of human parity in MT

made in 2018 have been addressed in the set-up of the experiments leading to the claims made in 2019, but that some of the issues still remain, we reassess these later claims.

The remainder of this paper is organised as follows. Section 2 discusses the potential issues in the setup of the human evaluation at WMT 2019. Next, in Section 3 we conduct a modified evaluation of the MT systems that reached human parity or super-human performance at WMT 2019. Finally, Section 4 presents our conclusions and recommendations.

2 Potential Issues in the Human Evaluation of WMT 2019

This section discusses the potential issues that we have identified in the human evaluation of the news translation task at WMT 2019, and motivates why they might have had contributed to the fact that some of the systems evaluated therein reached human parity or super-human performance. These issues concern the limited amount of intersentential context provided to the evaluators (Section 2.1), the fact that the evaluations were not conducted by translators (Section 2.2) and the fact that the evaluation was reference-based for one of the translation directions (Section 2.3).

2.1 Limited Intersentential Context

In the human evaluation at previous editions of WMT evaluators had no access to intersentential context since the sentences were shown to evaluators in random order. That changed in WMT 2019 (Barrault et al., 2019), which had two evaluation settings that contained intersentential context:

- Document-level (DR+DC), inspired by Läubli et al. (2018), in which the whole document is available and it is evaluated globally (see top of Figure 1). While the evaluator has access to the whole document, this set-up has the drawback of resulting in very few ratings (one per document) and hence suffers from low statistical power (Graham et al., 2019).
- Sentence-by-sentence with document context (SR+DC), in which segments are provided in the "natural order as they appear in the document" and they are assessed individually (see bottom of Figure 1). Such a set-up results in a much higher number of ratings compared to the previous evaluation setting (DR+DC): one per sentence rather than one per document. The problem with the current setting is that the evaluator can access limited intersentential context since only the current sentence is shown. This poses two issues, with respect to previous and following sentences in the document being evaluated. With respect to previous sentences, while the evaluator has seen them recently, he/she might have forgotten some details of a previous sentence that are relevant for the evaluation of the current sentence, e.g. in long documents. As for following sentences, the evaluator does not have access to them while evaluating the current sentence, which may be useful in some cases, e.g. when evaluating the first sentence of a document, i.e. the title of the newstory, since in some cases this may present an ambiguity for which having access to subsequent sentences could be useful.

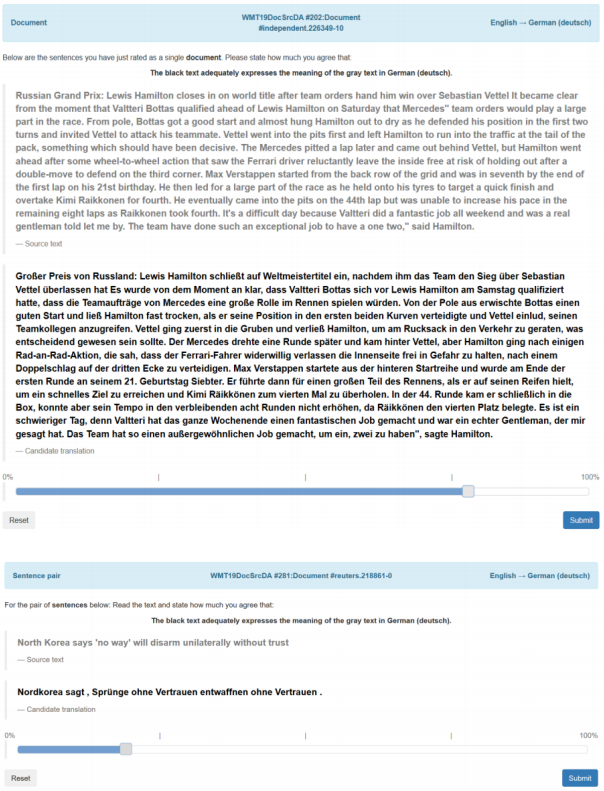

Figure 1: A snapshot of an assessment using setting DR+DC (top) and SR+DC (bottom) at WMT 2019, taken from Barrault et al. (2019)

SR+DC was the set-up used for the official rankings of WMT 2019, from which the claims of human parity and super-human performance were derived. The requirement of information from both previous and following sentences in human evalu-

ation of MT has been empirically proven in contemporary research (Castilho et al., in press 2020).

In our evaluation setup, evaluators are shown local context (the source sentences immediately preceding and following the current one) and are provided with global context: the whole source document as a separate text file. Evaluators are told to use the global context if the local context does not provide enough information to evaluate a sentence. In addition, evaluators are asked to evaluate all the sentences of a document in a single session.

2.2 Proficiency of the Evaluators

The human evaluation of WMT 2019 was conducted by crowd workers and by MT researchers. The first type of evaluators provided roughly two thirds of the judgments (487,674) while the second type contributed the remaining one third (242,424). Of the judgments provided by crowd workers, around half of them (224,046) were by "workers who passed quality control".

The fact that the evaluation was not conducted by translators might be problematic since it has been found that crowd workers lack knowledge of translation and, compared to professional translators, tend to be more accepting of (subtle) translation errors (Castilho et al., 2017).

Taking this into account, we will reassess the translations of the systems that achieved human parity or super-human performance at WMT 2019 with translators and non-translators. The latter are native speakers of the target language who are not translators but who have an advanced level of the source language (at least C1 in the Common European Framework of Reference for Languages).

2.3 Reference-based Evaluation

While for two of the translation directions for which there were claims of human parity at WMT 2019 the human evaluation was reference-free (from English to both German and Russian), for the remaining translation direction for which there was a claim of parity (German to English), the human evaluation was reference-based. In a reference-free evaluation, the evaluator assesses the quality of a translation with respect to the source sentence. Hence evaluators need to be proficient in both the source and target languages. Differently, in a reference-based evaluation, the evaluator assesses a translation with respect, not (only) to the source sentence, but (also) to a reference translation.

The advantage of a reference-based evaluation is that it can be carried out by monolingual speakers, since only proficiency in the target language is required. However, the dependence on reference translations in this type of evaluation can lead to reference bias. Such a bias is hypothesised to result in (i) inflated scores for candidate translations that happen to be similar to the reference translation (e.g. in terms of syntactic structure and lexical choice) and to (ii) penalise correct translations that diverge from the reference translation. Recent research has found both evicence that this is the case (Fomicheva and Specia, 2016; Bentivogli et al., 2018) and that it is not (Ma et al., 2017).

In the context of WMT 2019, in the translation directions that followed a reference-free human evaluation, the human translation (used as reference for the automatic evaluation) could be compared to MT systems in the human evaluation, just by being part of the pool of translations to be evaluated. However, in the translation directions that followed a reference-based human evaluation, such as German→English, the reference translation could not be evaluated against the MT systems, since it was itself the gold standard. A second human translation was used to this end. In a nutshell, for English→German and English→Russian there is one human translation, referred to as HUMAN, while for German→English there are two human translations, one was used as reference and the other was evaluated against the MT systems, to which we refer to as REF and HUMAN, respectively.

The claim of parity for German→English results therefore from the fact that HUMAN and the output of an MT system (Facebook-FAIR) were compared separately to a gold standard translation, REF, and the overall ratings that they obtained were not significantly different from each other. If there was reference bias in this case, it could be that HUMAN was penalised for being different than REF. To check whether this could be the case we use BLEU (Papineni et al., 2002) as a proxy to measure the similarity between all the pairs of the three relevant translations: REF, HUMAN and the best MT system. Table 1 shows the three pairwise scores.[1] HUMAN appears to be markedly differ-

[1] We use the `multi-bleu.perl` implementation of BLEU, giving as parameters one of the translations as the reference and the other as the hypothesis. Changing the order of the parameters results in very minor variations in the score.

ent than MT and REF, which are more similar to each other.

MT, REF	MT, HUMAN	REF, HUMAN
35.9	26.5	21.9

Table 1: BLEU scores between pairs of three translations (REF, HUMAN and the best MT system) for German→English at the news translation task of WMT 2019.

These results indicate thus that HUMAN could have been penalised for diverging from the reference translation REF, which could have contributed to the best MT system reaching parity. In our experiments, we will conduct a reference-free evaluation for this translation direction comparing this MT system to both human translations.

3 Evaluation

3.1 Experimental Setup

We conduct a human evaluation[2] for the three translation directions of WMT 2019 for which there were claims of human parity or super-human performance: German→English, English→German and English→Russian. We evaluate the first twenty documents of the test set for each of these language pairs. These amount to 317 sentences for German→English and 302 for both English→German and English→Russian (the English side of the test set in all from-English translation directions is common).

We conduct our evaluation with the Appraise toolkit (Federmann, 2012), by means of relative rankings, rather than direct assessment (DA) (Graham et al., 2017) as in Barrault et al. (2019). While DA has some advantages over ranking, their outcomes correlate strongly ($R > 0.9$ in Bojar et al. (2016)) and the latter is more appropriate for our evaluation for two reasons: (i) it allows us to show the evaluator all the translations that we evaluate at once, so that they are directly compared (DA only shows one translation at a time, entailing that the translations evaluated are indirectly compared to each other) and (ii) it allows us to show local context to the evaluator (DA only shows the sentence that is being currently evaluated).

Evaluators are shown two translations for both English→German and English→Russian: one by a human (referred to as HUMAN) and one by the best MT system[3] submitted to that translation direction (referred to as MT). For German→English there are three translations (see Section 2.3): two by humans (HUMAN and REF) and one by an MT system. The MT system is Facebook-FAIR for all three translation directions. The order in which the translations are shown is randomised.

For each source sentence, evaluators rank the translations thereof, with ties being allowed. Evaluators could also avoid ranking the translations of a sentence, if they detected an issue that prevented them from being able to rank them, by using the button flag error; they were instructed to do so only when strictly necessary. Figure 2 shows a snapshot of our evaluation.

From the relative rankings, we extract the number of times one of the translations is better than the other and the number of times they are tied. Statistical significance is conducted with two-tailed sign tests, the null hypothesis being that evaluators do not prefer the human translation over MT or viceversa (Läubli et al., 2018). We report the number of successes x, i.e. number of ratings in favour of the human translation, and the number of trials n, i.e. number of all ratings except for ties.

Five evaluators took part in the evaluation for English→German (two translators and three non-translators), six took part for English→Russian (four translators and two non-translators) and three took part for German→English (two translators and one non-translator).

Immediately after completing the evaluation, the evaluators completed a questionnaire (see Appendix A). It contained questions about their linguistic proficiency in the source and target languages, their amount of translation experience, the frequency with which they used the local and global contextual information, whether they thought that one of the translations was normally better than the other(s) and whether they thought that the translations were produced by human translators or MT systems.

In the remaining of this section we present the results of our evaluation for the three language pairs, followed by the inter-annotator agreement and the responses to the questionnaire.

[2]Code and data available at `https://github.com/antot/human_parity_eamt2020`

[3]The MT system with the highest normalised average DA score in the human evaluation of WMT 2019.

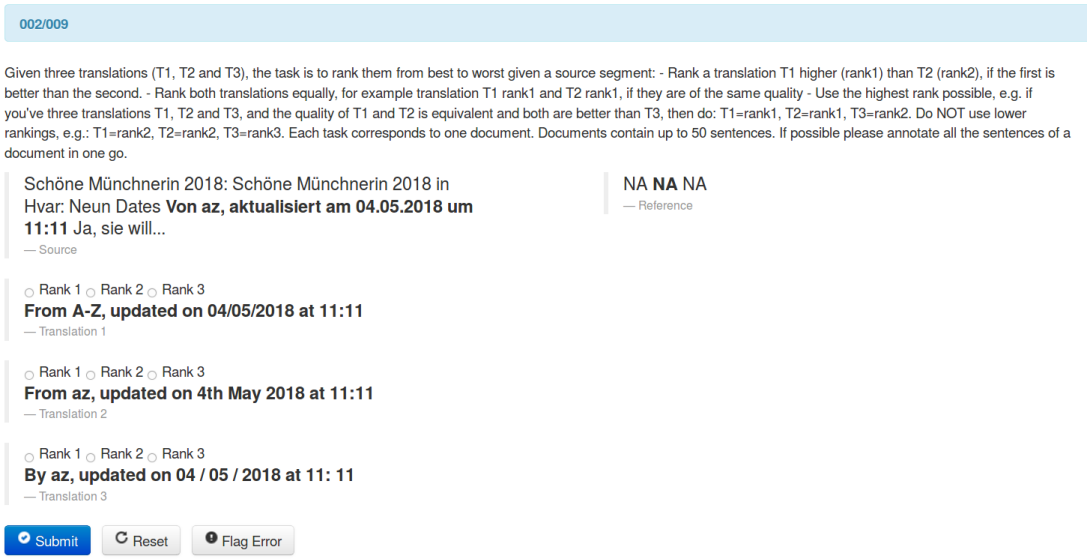

Figure 2: A snapshot of our human evaluation, for the German→English translation direction, for the second segment of a document that contains nine segments. The evaluator ranks three translations, two of which are produced by human translators (REF and HUMAN) while the remaining one comes from an MT system (Facebook-FAIR), by comparing them to the source, since no reference translation is provided. Local context (immediately preceeding and following source sentences) is provided inside the evaluation tool and global context (the whole source document) is provided as a separate file.

3.2 Results for English→German

Figure 3 shows the percentages of rankings[4] for which translators and non-translators preferred the translation by the MT system, that by the human translator or both were considered equivalent (tie). Non-translators preferred the translation by the MT engine slightly more frequently than the human translation (42.3% vs 36.7%) while the opposite is observed for translators (36.9% for HUMAN vs 34.9% for MT). However, these differences are not significant for either translators ($x = 222$, $n = 432$, $p = 0.6$) nor for non-translators ($x = 332$, $n = 715$, $p = 0.06$). In other words, according to our results there is no super-human performance, since MT is not found to be significantly better than HUMAN (which was the case at WMT 2019) but HUMAN is not significantly better than MT either. Therefore our evaluation results in human parity, since the performance of the MT system and HUMAN are not significantly different in the eyes of the translators and the non-translators that conducted the evaluation.

Figure 4 shows the results for each evaluator separately, with ties omitted to ease the visualisation. We observe a similar trend across all the non-translators: a slight preference for MT over

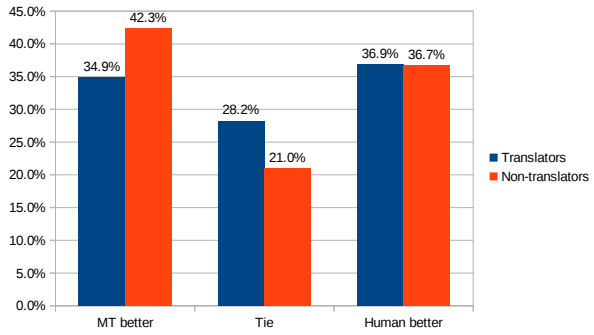

Figure 3: Results for English→German for translators ($n = 602$) and non-translators ($n = 905$)

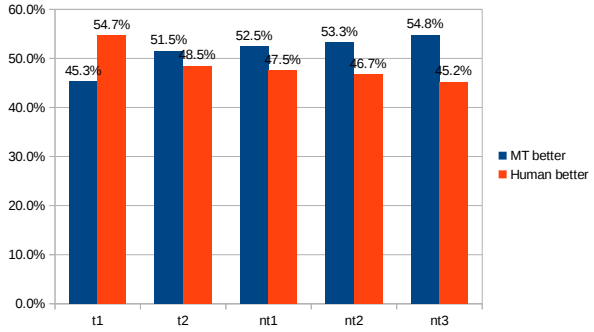

Figure 4: Results for English→German for each evaluator separately: translators t1 and t2 and non-translators nt1, nt2 and nt3.

[4]We show percentages instead of absolute numbers in order to be able to compare the rankings by translators and non-translators, as the number of translators and non-translators is not the same.

HUMAN, where the first is preferred in 52.5% to 54.8% of the times whereas the second is preferred in 45.2% to 47.5% of the cases. However, the two translators do not share the same trend; translator t1 prefers HUMAN more often than MT (54.7% vs 45.3%) while the trend is the opposite for translator t2, albeit more slightly (51.5% MT vs 48.5% HUMAN).

3.3 Results for English→Russian

Figure 5 shows the results for English→Russian. In this translation direction both translators and non-translators prefer HUMAN more frequently than MT: 42.3% vs 34.4% ($x = 499$, $n = 905$, $p < 0.01$) and 45.5% vs 35.8% ($x = 275$, $n = 491$, $p < 0.01$), respectively. Since the differences are significant in both cases, our evaluation refutes the claim of human parity made at WMT 2019 for this translation direction.

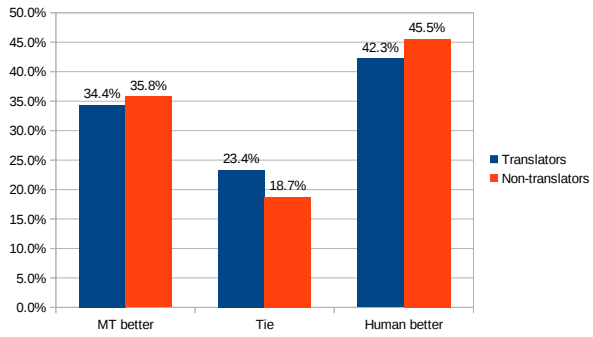

Figure 5: Results for English→Russian for translators ($n = 1181$) and non-translators ($n = 604$)

Again we zoom in on the results by the individual evaluators, as depicted in Figure 6. It can be seen that all but one of the evaluators, translator t1, prefer HUMAN considerably more often than MT. However, the differences are only significant for t3 ($x = 114$, $n = 178$, $p < 0.001$) and nt2 ($x = 119$, $n = 202$, $p < 0.05$), probably due to the small number of observations.

3.4 Results for German→English

As explained in section 2.3, for this translation direction there are two human translations, referred to as HUMAN and REF, and one MT system. Hence we can establish three pairwise comparisons: REF–MT, HUMAN–MT and HUMAN–REF. The results for them are shown in Figure 7, Figure 8 and Figure 9, respectively.

Both translators preferred the translation by the MT system slightly more often than the human

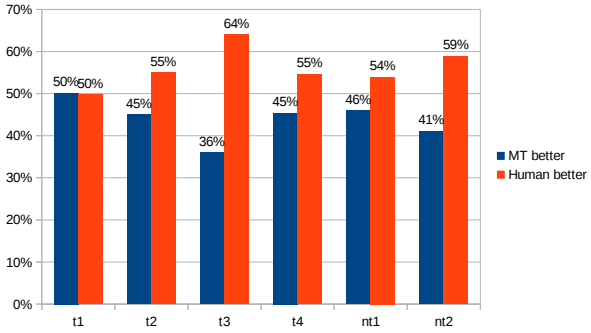

Figure 6: Results for English→Russian for each evaluator separately: translators t1, t2, t3 and t4 and non-translators nt1 and nt2.

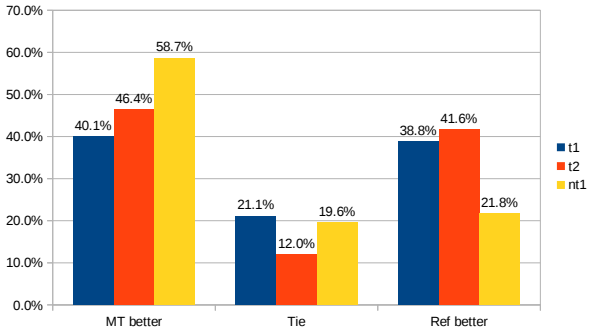

Figure 7: Results for German→English for REF and MT, with translators t1 and t2 and non-translator nt1.

translation REF, 40% vs 39% and 46% vs 42%, but the difference is not significant ($x = 255$, $n = 529$, $p = 0.4$). The non-translator preferred the translation by MT considerably more often than REF: 59% vs 22%, with the diffence being significant ($x = 69$, $n = 255$, $p < 0.001$). In other words, compared to REF, the human translation used as gold standard at WMT 2019, the MT system achieves human parity according to the two translators and super-human performance according to the non-translator.

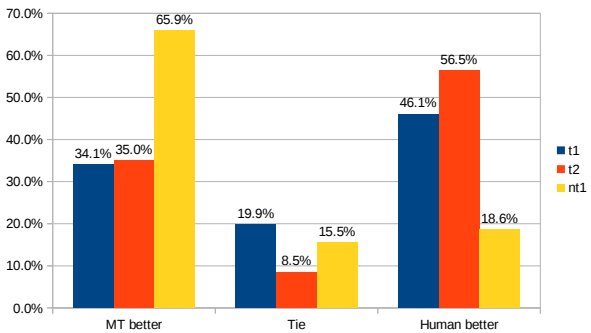

Figure 8: Results for German→English for HUMAN and MT, with translators t1 and t2 and non-translator nt1.

Now we discuss the results of comparing the

MT system to the other human translation, HUMAN (see Figure 8). The outcome according to the non-translator is, as in the previous comparison between REF and MT, super-human performance ($x = 59$, $n = 268$, $p < 0.001$), which can be expected since this evaluator prefers MT much more often than HUMAN: 66% vs 19% of the times. We expected that the results for the translators would also follow a similar trend to their outcome when they compared MT to the other human translation (REF), i.e. human parity. However, we observe a clear preference for HUMAN over MT: 46% vs 34% and 57% vs 35%, resulting in a significant difference ($x = 325$, $n = 544$, $p < 0.001$).

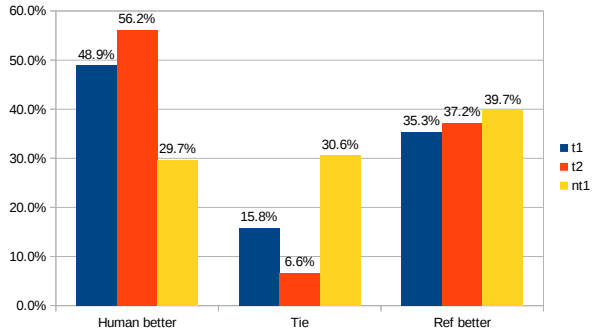

Figure 9: Results for German→English for REF and HUMAN, with translators t1 and t2 and non-translator nt1.

The last comparison is shown in Figure 9 and concerns the two human translations: REF and HUMAN. The two translators exhibit a clear preference for HUMAN over REF: 49% vs 35% and 56% vs 37%, ($x = 230$, $n = 563$, $p < 0.001$). Conversely, the non-translator preferred REF significantly more often than HUMAN ($x = 126$, $n = 220$, $p < 0.05$): 40% vs 30%.

Given that (i) parity was found between MT and HUMAN in the reference-based evaluation of WMT, where REF was the reference translation, that (ii) HUMAN is considerably different than REF and MT (see Section 2.3) and that (iii) HUMAN is found to be significantly better than REF by translators in our evaluation, it seems that reference bias played a role in the claim of parity at WMT.

3.5 Results of the Inter-annotator Agreement

We now report the inter-annotator agreement (IAA) between the evaluators. Since we have two types of evaluators, translators and non-translators, we report the IAA for both of them. IAA is calculated in terms of Cohen's kappa coefficient (κ) as it was done at WMT 2016 (Bojar et al., 2016, Section 3.3).

Direction	Evaluators	
	ts	nts
English→German	0.326	0.266
English→Russian	0.239	0.238
German→English	0.320	NA

Table 2: Inter-annotator agreement with Cohen's κ among translators (ts) and non-translators (nts) for the three translation directions.

Table 2 shows the IAA coefficients. For English→German, the IAA among translators ($\kappa = 0.326$) is considerably higher, 23% relative, than among non-translators ($\kappa = 0.266$). For English→Russian, both types of evaluators agree at a very similar level ($\kappa = 0.239$ and $\kappa = 0.238$). Finally, for German→English, we cannot establish a direct comparison between the IAA of translators and non-translators, since there was only one non-translator. However, we can compare the IAA of the two translators ($\kappa = 0.32$) to that of each of the translators and the non-translator: $\kappa = 0.107$ between the first translator and the non-translator and $\kappa = 0.125$ between the second translator and the non-translator. The agreement between translators is therefore 176% higher than between one translator and the non-translator.

In a nutshell, for the three translation directions the IAA of translators is higher than, or equivalent to, that of non-translators, which corroborates previous findings by Toral et al. (2018), where the IAA was 0.254 for translators and 0.13 for non-translators.

3.6 Results of the Questionnaire

The questionnaire (see Appendix A) contained two 5-point Likert questions about how often additional context, local and global, was used. In both cases, translators made slightly less use of context than non-translators: $M = 2.9$, $SD = 2.0$ versus $M = 3.5$, $SD = 1.0$ for local context and $M = 1.4$, $SD = 0.7$ versus $M = 2$, $SD = 0.9$ for global context. Our interpretation is that translators felt more confident to rank the translations and thus used additional contextual information to a lesser extent. If an evaluator used global context, they were asked to specify whether they used it mostly for some sentences in particular (those at the beginning, middle or at the end of the documents) or not. Out of 8 respondents, 5 reported to have used global context mostly for sentences re-

gardless of their position in the document and the remaining 3 mostly for sentences at the beginning.

In terms of the perceived quality of the translations evaluated, all non-translators found one of the translations to be clearly better in general. Five out of the eight translators gave that reply too while the other three translators found all translations to be of similar quality (not so good).

Asked whether they thought the translations had been produced by MT systems or by humans, all evaluators replied that some were by humans and some by MT systems, except one translator, who thought that all the translations were by MT systems, and one non-translator who answered that he/she did not know.

4 Conclusions and Future Work

We have conducted a modified evaluation on the MT systems that reached human parity or super-human performance at the news shared task of WMT 2019. According to our results: (i) for English→German, the claim of super-human performance is refuted, but there is human parity; (ii) for English→Russian, the claim of human parity is refuted; (iii) for German→English, for which there were two human translations, the claim of human parity is refuted with respect to the best of the human translations, but not with respect to the worst.

Based on our findings, we put forward a set of recommendations for human evaluation of MT in general and for the assessment of human parity in MT in particular:

1. Global context (i.e. the whole document) should be available to the evaluator. Some of the evaluators have reported that they needed that information to conduct some of the rankings and contemporary research (Castilho et al., in press 2020) has demonstrated that such knowledge is indeed required for the evaluation of some sentences.
2. If the evaluation is to be as accurate as possible then it should be conducted by professional translators. Our evaluation has corroborated that evaluators that do not have translation proficiency evaluate MT systems more leniently than translators and that inter-annotator agreement is higher among the latter (Toral et al., 2018).
3. Reference-based human evaluation should be in principle avoided, given the reference bias issue (Bentivogli et al., 2018), which according to our results seems to have played a role in the claim of human parity for German→English at WMT 2019. That said, we note that there is also research that concludes that there is no evidence of reference bias (Ma et al., 2017).

The first two recommendations were put forward recently (Läubli et al., 2020) and are corroborated by our findings. We acknowledge that our conclusions and recommendations are somewhat limited since they are based on a small number of sentences (just over 300 for each translation direction) and evaluators (14 in total).

Claims of human parity are of course not specific to translation. Super-human performance has been reported to have been achieved in many other tasks, including board games, e.g. chess (Hsu, 2002) and Go (Silver et al., 2017). However, we argue that assessing human parity in translation, and probably in other language-related tasks too, is not as straightforward as in other tasks such as board games, and that the former task poses, at least, two open questions, which we explore briefly in the following to close the paper.

1. Against whom should the machine be evaluated? In other words, should one claim human parity if the output of an MT system is perceived to be indistiguishable from that by an *average* professional translator or should we only compare to a *champion* professional translator? In other tasks it is the latter case, e.g. chess in which DEEP BLUE outperformed world champion Gary Kasparov. Related, we note that in tasks such as chess it is straightforward to define the concept of a player being better than another: whoever wins more games, the rules of which are deterministic. But in the case of translation, it is not so straightforward to define whether a translator is better than another. This question is pertinent since, as we have seen for German→English (Section 3.4), where we had translations by two professional translators, the choice of which one is used to evaluate an MT system against can lead to a claim of human parity or not. In addition, the reason why one claim remains after our evaluation (human parity for English→German) might be that the human translation therein is not *as*

good as it could be. Therefore, once the three potential issues that we have put forward (see Section 2) are solved, we think that an important potential issue that should be studied, and which we have not considered, has to do with the quality of the human translation used.

2. Who should assess claims of human parity and super-human performance? Taking again the example of chess, this is straightforward since one can just count how many games each contestant (machine and human) wins. In translation, however, we need a person with knowledge of both languages to assess the translations. We have seen that the outcome is dependent to some extent on the level of translation proficiency of the evaluator: it is more difficult to find human parity if the translations are evaluated by professional translators than if the evaluation is carried out by bilingual speakers without any translation proficiency. Taking into account that most of the users of MT systems are not translators, should we in practice consider human parity if those users do not perceive a significant difference between human and machine translations, even if an experienced professional translator does?

Acknowledgments

This research has received funding from CLCG's 2019 budget for research participants. I am grateful for valuable comments from Barry Haddow, co-organiser of WMT 2019. I would also like to thank the reviewers; their comments have definitely led to improve this paper.

References

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

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas.

Bentivogli, Luisa, Mauro Cettolo, Marcello Federico, and Federmann Christian. 2018. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *15th International Workshop on Spoken Language Translation 2018*, pages 62–69.

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

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

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, and Maria Gialama. 2017. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *MT Summit 2017*, pages 116–131, Nagoya, Japan.

Castilho, Sheila, Maja Popovic, and Andy Way. in press 2020. On Context Span Needed for Machine Translation Evaluation. In *Proceedings of the Twelvfth International Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA).

Federmann, Christian. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35, September.

Fomicheva, Marina and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 77–82, Berlin, Germany, August. Association for Computational Linguistics.

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

Graham, Yvette, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*.

Haddow, Barry, Nikolay Bogoychev, Denis Emelin, Ulrich Germann, Roman Grundkiewicz, Kenneth Heafield, Antonio Valerio Miceli Barone, and Rico Sennrich. 2018. The university of edinburgh's submissions to the wmt18 news translation task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 403–413, Belgium, Brussels, October. Association for Computational Linguistics.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation.

Hsu, Feng-Hsiung. 2002. *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*. Princeton University Press, Princeton, NJ, USA.

Kocmi, Tom, Roman Sudarikov, and Ondej Bojar. 2018. Cuni submissions in wmt18. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 435–441, Belgium, Brussels, October. Association for Computational Linguistics.

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

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

Ng, Nathan, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fairs wmt19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy, August. Association for Computational Linguistics.

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

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *CoRR*, abs/1712.01815.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 113–123, Belgium, Brussels, October. Association for Computational Linguistics.

A Post-experiment Questionnaire

1. Rate your knowledge of the source language
   - None; A1; A2; B1; B2; C1; C2; native
2. Rate your knowledge of the target language
   - None; A1; A2; B1; B2; C1; C2; native
3. How much experience do you have translating from the source to the target language?
   - None, and I am not a translator; None, but I am a translator; Less than 1 year; between 1 and 2 years; between 2 and 5 years; more than 5 years
4. During the experiment, how often did you use the local context shown in the web application (i.e. source sentences immediately preceding and immediately following the current sentence)?
   - Never; rarely; sometimes; often; always
5. During the experiment, how often did you use the global context provided (i.e. the whole source document provided as a text file)?
   - Never; rarely; sometimes; often; always
6. If you used the global context, was that the case for ranking some sentences in particular?
   - Yes, mainly those at the beginning of documents, e.g. headlines
   - Yes, mainly those in the middle of documents
   - Yes, mainly those at the end of documents
   - No, I used the global context regardless of the position of the sentences to be ranked
7. About the translations you ranked
   - Normally one was clearly better
   - All were of similar quality, and they were not so good
   - All were of similar quality, and they were very good
8. The translations that you evaluated were in your opinion:
   - All produced by human translators
   - All produced by machine translation systems
   - Some produced by humans and some by machine translation systems
   - I don't know