Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, March : 2023

# ASSESSMENT OF HUMAN-MACHINE PARITY IN 'LANGUAGE TRANSLATION'

**JANGAM SWAPNA,** Assistant Professor, V.K.R, V.N.B & A.G.K. College of Engineering, GUDIVADA Krishna Dist. Andhra Pradesh
**MAHALI PRASANTHI,** Associate Professor, V.K.R, V.N.B & A.G.K. College of Engineering, GUDIVADA Krishna Dist. Andhra Pradesh

**Abstract**
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. We find that for all three aspects, human translations are judged more favorably, 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.

**Key Words:** analysis, data, evaluation, language, linguistic, methods, research, translation

## 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.

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).

## 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 minimize 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, butthe studies supporting the use of crowd sourcing 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 developersto which crowd workers were compared are usually not professional translators. We hypothesize 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 sentence level, owing to the fact that most MT systems do not yet take context across sentence boundariesinto 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 qualityless favourably when evaluating full documents.

**Reference Translations** The human reference translations with which machine transla- tions 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 halfwere 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 sometimesshow 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 ex- ample, had the WMT source texts re-translated as they were not convinced of the qualityof the human translations in the test set. At WMT 2018, the organizers themselves noted that "the manual evaluation included several reports of ill-formed reference

translations" (Bojar et al., 2018, p. 292). We hypothesize that the quality of the human translationshas a significant effect on findings of human–machine parity, which would indicate thatit is necessary to ensure that human translations used to assess parity claims need to be carefully vetted for their quality.

## 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:

**$H_A$** The professional human translations in the dataset of Hassan et al. (2018).

**$H_B$** Professional human translations that we ordered from a different translation vendor, which included a post-hoc native English check.

**$MT_1$** The machine translations produced by Hassan et al.'s (2018) best system (COMBO-6), 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.

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

## 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 thereis 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 con-fronted with a translation that is hard to post-edit, tend to accept the MT rather than tryto improve it.

## 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.

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 testset originally written in Chinese, and are shown in their original order. Raters are shownone 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 ex- perts were professional Chinese to English translators: one native in Chinese with a fluent 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, i. e., we run 1,000 iterations of the rankings recorded with Appraise followed by clustering (significance level $\alpha = 0.05$).

| Rank | Translators | | | | | |
|---|---|---|---|---|---|---|
| | All | | Experts | | Non-experts | |
| | $n = 3873$ | | $n = 1785$ | | $n = 2088$ | |
| 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 $p \leq .05$.

## 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).

## 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 rela- tively 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).

**Evaluation Protocol**

We test if the availability of document-level context affects human–machine parity claimsin 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 otherby 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 testset 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 itemas better than or equal to an actual translation, this is a strong indication that they didnot 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 and 104 sentences.

| 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 $MT_1$ islower 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.

**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 lan- guage, they may make mistakes in interpreting the original message. This is partic-ularly 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 doinga 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, Vande- pitte, Hartsuiker, & Macken, 2017; Toral, 2019).

## 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) foundno significant difference between the two, concluding that their NMT system had reached parity with professional human translation. Our blind qualitative analysis, however, showedthat 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 foundthat 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, whilenon-expert raters did not. In terms of linguistic context, raters found human translation significantly more accurate than machine translation when evaluating full documents, but 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). 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, whichwe 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 validityof 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

1. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by JointlyLearning to Align and Translate. In *Proceedings of ICLR*. San Diego, CA.
2. Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y.,

3. Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus Phrase-BasedMachine Translation Quality: a Case Study. In *Proceedings of EMNLP* (pp. 257–267).Austin, Texas.
4. 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).
5. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., . . . Zampieri,
6. C. (2018). Findings of the 2018 Conference on Machine Translation (WMT18). In
7. *Proceedings of WMT* (pp. 272–307). Belgium, Brussels.
8. Callison-Burch, C. (2009). Fast, Cheap, and Creative: Evaluating Translation Quality UsingAmazon's Mechanical Turk. In *Proceedings of EMNLP* (pp. 286–295). Singapore.
9. 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.
10. 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.
11. 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.
12. 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, Bul- garia.
13. Daems, J., De Clercq, O., & Macken, L. (2017). Translationese and Post-editese: How comparable is comparable quality? *Linguistica Antverpiensia, New Series – Themesin Translation Studies*, *16* , 89–103.
14. 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.
15. Doherty, S. (2017). Issues in human and automatic translation quality assessment. In *Human issues in translation technology* (pp. 131–148). Routledge.
16. 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.
17. 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.)
18. 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.
19. 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.
20. 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.
21. Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., . . . Zhou, M. (2018). Achieving Human Parity on Automatic Chinese to English News Transla- tion. *arXiv preprint 1803.05567* .

22. Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of CHI* (pp. 453–456). Florence, Italy.
23. 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).
24. Laviosa, S. (1998). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta*, *43* (4), 557–570.
25. Laviosa-Braithwaite, S. (1998). Universals of translation. In *Routledge Encyclopedia of Translation Studies* (pp. 288–291). Routledge.
26. 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.
27. Neubig, G., Morishita, M., & Nakamura, S. (2015). Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015. In *Proceedings of WAT2015*. Kyoto, Japan.
28. Popel, M. (2018). CUNI Transformer Neural MT System for WMT18. In *Proceedings of WMT* (pp. 486–491). Brussels, Belgium.
29. 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.
30. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Proceedings of NIPS* (pp. 3104–3112). Montreal, Canada.
31. Toral, A. (2019). Post-editese: an Exacerbated Translationese. In *Proceedings of MT Summit* (pp. 273–281). Dublin, Ireland: European Association for Machine Translation.
32. 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.
33. Vilar, D., Xu, J., Luis Fernando, D., & Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of LREC* (pp. 697–702).
34. 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.
35. 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.