Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation

Samuel Läubli[1] **Rico Sennrich**[1,2] **Martin Volk**[1]

[1]Institute of Computational Linguistics, University of Zurich
{laeubli,volk}@cl.uzh.ch

[2]School of Informatics, University of Edinburgh
rico.sennrich@ed.ac.uk

Abstract

Recent research suggests that neural machine translation achieves parity with professional human translation on the WMT Chinese–English news translation task. We empirically test this claim with alternative evaluation protocols, contrasting the evaluation of single sentences and entire documents. In a pairwise ranking experiment, human raters assessing adequacy and fluency show a stronger preference for human over machine translation when evaluating documents as compared to isolated sentences. Our findings emphasise the need to shift towards document-level evaluation as machine translation improves to the degree that errors which are hard or impossible to spot at the sentence-level become decisive in discriminating quality of different translation outputs.

1 Introduction

Neural machine translation (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015) has become the de-facto standard in machine translation, outperforming earlier phrase-based approaches in many data settings and shared translation tasks (Luong and Manning, 2015; Sennrich et al., 2016; Cromieres et al., 2016). Some recent results suggest that neural machine translation "approaches the accuracy achieved by average bilingual human translators [on some test sets]" (Wu et al., 2016), or even that its "translation quality is at human parity when compared to professional human translators" (Hassan et al., 2018). Claims of human parity in machine translation are certainly extraordinary, and require extraordinary evidence.[1] Laudably, Hassan et al. (2018) have released their data publicly to allow external validation of their claims. Their claims are further strengthened by the fact that they follow best practices in human machine translation evaluation, using evaluation protocols and tools that are also used at the yearly Conference on Machine Translation (WMT) (Bojar et al., 2017), and take great care in guarding against some confounds such as test set selection and rater inconsistency.

However, the implications of a statistical tie between two machine translation systems in a shared translation task are less severe than that of a statistical tie between a machine translation system and a professional human translator, so we consider the results worthy of further scrutiny. We perform an independent evaluation of the professional translation and best machine translation system that were found to be of equal quality by Hassan et al. (2018). Our main interest lies in the evaluation protocol, and we empirically investigate if the lack of document-level context could explain the inability of human raters to find a quality difference between human and machine translations. We test the following hypothesis:

> A professional translator who is asked to rank the quality of two candidate translations on the document level will prefer a professional human translation over a machine translation.

Note that our hypothesis is slightly different from that tested by Hassan et al. (2018), which could be phrased as follows:

> A bilingual crowd worker who is asked to directly assess the quality of candidate translations on the sentence level will prefer a professional human translation over a machine translation.

[1]The term "parity" may raise the expectation that there is evidence for equivalence, but the term is used in the definition of "there [being] no statistical significance between [two outputs] for a test set of candidate translations" by Hassan et al. (2018). Still, we consider this finding noteworthy given the strong evaluation setup.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4791–4796
Brussels, Belgium, October 31 - November 4, 2018.

As such, our evaluation is not a direct replication of that by Hassan et al. (2018), and a failure to reproduce their findings does not imply an error on either our or their part. Rather, we hope to indirectly assess the accuracy of different evaluation protocols. Our underlying assumption is that professional human translation is still superior to neural machine translation, but that the sensitivity of human raters to these quality differences depends on the evaluation protocol.

2 Human Evaluation of Machine Translation

Machine translation is typically evaluated by comparing system outputs to source texts, reference translations, other system outputs, or a combination thereof (for examples, see Bojar et al., 2016a). The scientific community concentrates on two aspects: adequacy, typically assessed by bilinguals; and target language fluency, typically assessed by monolinguals. Evaluation protocols have been subject to controversy for decades (e. g., Van Slype, 1979), and we identify three aspects with particular relevance to assessing human parity: granularity of measurement (ordinal vs. interval scales), raters (experts vs. crowd workers), and experimental unit (sentence vs. document).

2.1 Related Work

Granularity of Measurement Callison-Burch et al. (2007) show that ranking (*Which of these translations is better?*) leads to better inter-rater agreement than absolute judgement on 5-point Likert scales (*How good is this translation?*) but gives no insight about how much a candidate translation differs from a (presumably perfect) reference. To this end, Graham et al. (2013) suggest the use of continuous scales for direct assessment of translation quality. Implemented as a slider between 0 (*Not at all*) and 100 (*Perfectly*), their method yields scores on a 100-point interval scale in practice (Bojar et al., 2016b, 2017), with each raters' rating being standardised to increase homogeneity. Hassan et al. (2018) use source-based direct assessment to avoid bias towards reference translations. In the shared task evaluation by Cettolo et al. (2017), raters are shown the source and a candidate text, and asked: *How accurately does the above candidate text convey the semantics of the source text?* In doing so, they have translations produced by humans and machines rated independently, and parity is assumed if the mean score of the former does not significantly differ from the mean score of the latter.

Raters To optimise cost, machine translation quality is typically assessed by means of crowdsourcing. Combined ratings of bilingual crowd workers have been shown to be more reliable than automatic metrics and "very similar" to ratings produced by "experts"[2] (Callison-Burch, 2009). Graham et al. (2017) compare crowdsourced to "expert" ratings on machine translations from WMT 2012, concluding that, with proper quality control, "machine translation systems can indeed be evaluated by the crowd alone." However, it is unclear whether this finding carries over to translations produced by NMT systems where, due to increased fluency, errors are more difficult to identify (Castilho et al., 2017a), and concurrent work by Toral et al. (2018) highlights the importance of expert translators for MT evaluation.

Experimental Unit Machine translation evaluation is predominantly performed on single sentences, presented to raters in random order (e. g., Bojar et al., 2017; Cettolo et al., 2017). There are two main reasons for this. The first is cost: if raters assess entire documents, obtaining the same number of data points in an evaluation campaign multiplies the cost by the average number of sentences per document. The second is experimental validity. When comparing systems that produce sentences without considering document-level context, the perceived suprasentential cohesion of a system output is likely due to randomness and thus a confounding factor. While incorporating document-level context into machine translation systems is an active field of research (Webber et al., 2017), state-of-the-art systems still operate at the level of single sentences (Sennrich et al., 2017; Vaswani et al., 2017; Hassan et al., 2018). In contrast, human translators can and do take document-level context into account (Krings, 1986). The same holds for raters in evaluation campaigns. In the discussion of their results, Wu et al. (2016) note that their raters "[did] not necessarily fully understand each randomly sampled sentence sufficiently" because it was provided with no context. In such setups, raters cannot reward textual cohesion and coherence.

[2] "Experts" here are computational linguists who develop MT systems, who may not be expert translators.

2.2 Our Evaluation Protocol

We conduct a quality evaluation experiment with a 2×2 mixed factorial design, testing the effect of source text availability (adequacy, fluency) and experimental unit (sentence, document) on ratings by professional translators.

Granularity of Measurement We elicit judgements by means of pairwise ranking. Raters choose the better (with ties allowed) of two translations for each item: one produced by a professional translator (HUMAN), the other by machine translation (MT). Since our evaluation includes that of human translation, it is reference-free. We evaluate in two conditions: adequacy, where raters see source texts and translations (*Which translation expresses the meaning of the source text more adequately?*); and fluency, where raters only see translations (*Which text is better English?*).

Raters We recruit professional translators, only considering individuals with at least three years of professional experience and positive client reviews.

Experimental Unit To test the effect of context on perceived translation quality, raters evaluate entire documents as well as single sentences in random order (i. e., context is a within-subjects factor). They are shown both translations (HUMAN and MT) for each unit; the source text is only shown in the adequacy condition.

Quality Control To hedge against random ratings, we convert 5 documents and 16 sentences per set into spam items (Kittur et al., 2008): we render one of the two options nonsensical by shuffling its words randomly, except for 10 % at the beginning and end.

Statistical Analysis We test for statistically significant preference of HUMAN over MT or vice versa by means of two-sided Sign Tests. Let a be the number of ratings in favour of MT, b the number of ratings in favour of HUMAN, and t the number of ties. We report the number of successes x and the number of trials n for each test, such that $x = b$ and $n = a + b$.[3]

[3] Emerson and Simon (1979) suggest the inclusion of ties such that $x = b + 0.5t$ and $n = a + b + t$. This modification has no effect on the significance levels reported in this paper.

2.3 Data Collection

We use the experimental protocol described in the previous section for a quality assessment of Chinese to English translations of news articles. To this end, we randomly sampled 55 documents and 2×120 sentences from the WMT 2017 test set. We only considered the 123 articles (documents) which are native Chinese,[4] containing 8.13 sentences on average. Human and machine translations (REFERENCE-HT as HUMAN, and COMBO-6 as MT) were obtained from data released by Hassan et al. (2018).[5]

The sampled documents and sentences were rated by professional translators we recruited from ProZ:[6] 4 native in Chinese (2), English (1), or both (1) to rate adequacy, and 4 native in English to rate fluency. On average, translators had 13.7 years of experience and 8.8 positive client reviews on ProZ, and received US$ 188.75 for rating 55 documents and 120 sentences.

The averages reported above include an additional translator we recruited when one rater showed poor performance on document-level spam items in the fluency condition, whose judgements we exclude from analysis. We also exclude sentence-level results from 4 raters because there was overlap with the documents they annotated, which means that we cannot rule out that the sentence-level decisions were informed by access to the full document. To allow for external validation and further experimentation, we make all experimental data publicly available.[7]

3 Results

In the adequacy condition, MT and HUMAN are not statistically significantly different on the sentence level ($x = 86$, $n = 189$, $p = .244$). This is consistent with the results Hassan et al. (2018) obtained with an alternative evaluation protocol (crowdsourcing and direct assessment; see Section 2.1). However, when evaluating entire doc-

[4] While it is common practice in machine translation to use the same test set in both translation directions, we consider a direct comparison between human "translation" and machine translation hard to interpret if one is in fact the original English text, and the other an automatic translation into English of a human translation into Chinese. In concurrent work, Toral et al. (2018) expand on the confounding effect of evaluating text where the target side is actually the original document.

[5] http://aka.ms/Translator-HumanParityData

[6] https://www.proz.com

[7] https://github.com/laeubli/parity

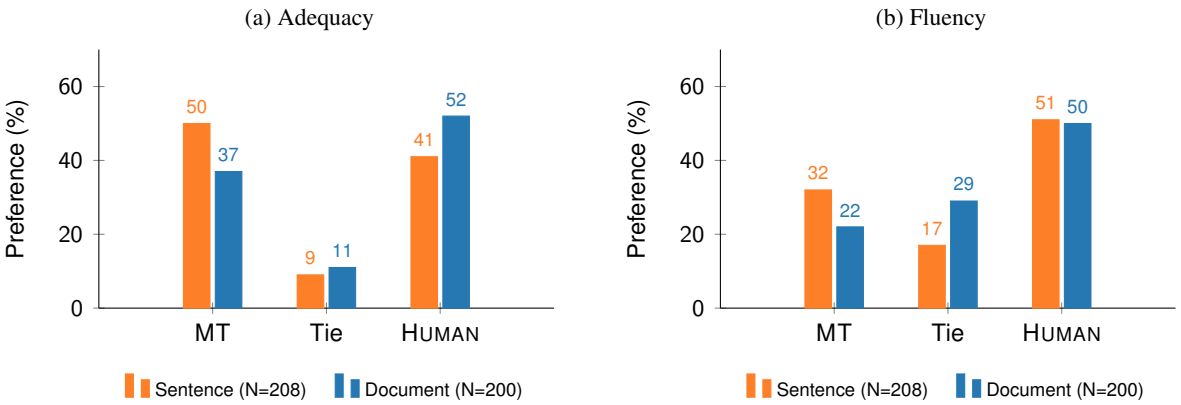

Figure 1: Raters prefer human translation more strongly in entire documents. When evaluating isolated sentences in terms of adequacy, there is no statistically significant difference between HUMAN and MT; in all other settings, raters show a statistically significant preference for HUMAN.

uments, raters show a statistically significant preference for HUMAN ($x = 104$, $n = 178$, $p < .05$). While the number of ties is similar in sentence- and document-level evaluation, preference for MT drops from 50 to 37 % in the latter (Figure 1a).

In the fluency condition, raters prefer HUMAN on both the sentence ($x = 106$, $n = 172$, $p < .01$) and document level ($x = 99$, $n = 143$, $p < .001$). In contrast to adequacy, fluency ratings in favour of HUMAN are similar in sentence- and document-level evaluation, but raters find more ties with document-level context as preference for MT drops from 32 to 22 % (Figure 1b).

We note that these large effect sizes lead to statistical significance despite modest sample size. Inter-annotator agreement (Cohen's κ) ranges from 0.13 to 0.32 (see Appendix for full results and discussion).

4 Discussion

Our results emphasise the need for suprasentential context in human evaluation of machine translation. Starting with Hassan et al.'s (2018) finding of no statistically significant difference in translation quality between HUMAN and MT for their Chinese–English test set, we set out to test this result with an alternative evaluation protocol which we expected to strengthen the ability of raters to judge translation quality. We employed professional translators instead of crowd workers, and pairwise ranking instead of direct assessment, but in a sentence-level evaluation of adequacy, raters still found it hard to discriminate between HUMAN and MT: they did not show a statistically significant preference for either of them.

Conversely, we observe a tendency to rate HUMAN more favourably on the document level than on the sentence level, even within single raters. Adequacy raters show a statistically significant preference for HUMAN when evaluating entire documents. We hypothesise that document-level evaluation unveils errors such as mistranslation of an ambiguous word, or errors related to textual cohesion and coherence, which remain hard or impossible to spot in a sentence-level evaluation. For a subset of articles, we elicited both sentence-level and document-level judgements, and inspected articles for which sentence-level judgements were mixed, but where HUMAN was strongly preferred in document-level evaluation. In these articles, we do indeed observe the hypothesised phenomena. We find an example of lexical coherence in a 6-sentence article about a new app "微信挪车", which HUMAN consistently translates into "WeChat Move the Car". In MT, we find three different translations in the same article: "Twitter Move Car", "WeChat mobile", and "WeChat Move". Other observations include the use of more appropriate discourse connectives in HUMAN, a more detailed investigation of which we leave to future work.

To our surprise, fluency raters show a stronger preference for HUMAN than adequacy raters (Figure 1). The main strength of neural machine translation in comparison to previous statistical approaches was found to be increased fluency, while adequacy improvements were less clear (Bojar et al., 2016b; Castilho et al., 2017b), and we expected a similar pattern in our evaluation. Does this indicate that adequacy is in fact a strength of

MT, not fluency? We are wary to jump to this conclusion. An alternative interpretation is that MT, which tends to be more literal than HUMAN, is judged more favourably by raters in the bilingual condition, where the majority of raters are native speakers of the source language, because of L1 interference. We note that the availability of document-level context still has a strong impact in the fluency condition (Section 3).

5 Conclusions

In response to recent claims of parity between human and machine translation, we have empirically tested the impact of sentence and document level context on human assessment of machine translation. Raters showed a markedly stronger preference for human translations when evaluating at the level of documents, as compared to an evaluation of single, isolated sentences.

We believe that our findings have several implications for machine translation research. Most importantly, if we accept our interpretation that human translation is indeed of higher quality in the dataset we tested, this points to a failure of current best practices in machine translation evaluation. As machine translation quality improves, translations will become harder to discriminate in terms of quality, and it may be time to shift towards document-level evaluation, which gives raters more context to understand the original text and its translation, and also exposes translation errors related to discourse phenomena which remain invisible in a sentence-level evaluation.

Our evaluation protocol was designed with the aim of providing maximal validity, which is why we chose to use professional translators and pairwise ranking. For future work, it would be of high practical relevance to test whether we can also elicit accurate quality judgements on the document-level via crowdsourcing and direct assessment, or via alternative evaluation protocols. The data released by Hassan et al. (2018) could serve as a test bed to this end.

One reason why document-level evaluation widens the quality gap between machine translation and human translation is that the machine translation system we tested still operates on the sentence level, ignoring wider context. It will be interesting to explore to what extent existing and future techniques for document-level machine translation can narrow this gap. We expect that this will require further efforts in creating document-level training data, designing appropriate models, and supporting research with discourse-aware automatic metrics.

Acknowledgements

We thank Xin Sennrich for her help with the analysis of translation errors. We also thank Antonio Toral and the anonymous reviewers for their helpful comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*, San Diego, CA.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016a. Ten years of WMT evaluation campaigns: Lessons learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*, pages 27–34, Portorož, Slovenia.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of WMT*, pages 169–214, Copenhagen, Denmark.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016b. Findings of the 2016 Conference on Machine Translation (WMT16). In *Proceedings of WMT*, pages 131–198, Berlin, Germany.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of EMNLP*, pages 286–295, Singapore.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of WMT*, pages 136–158, Prague, Czech Republic.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017a. Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, 108:109–120.

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

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuitho Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of IWSLT*, pages 2–14, Tokyo, Japan.

Fabien Cromieres, Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2016. Kyoto University participation to WAT 2016. In *Proceedings of WAT*, pages 166–174, Osaka, Japan.

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria.

Yvette Graham, 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.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William 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. *Computing Research Repository*, arXiv:1803.05567.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of EMNLP*, pages 1700–1709, Seattle, WA.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of CHI*, pages 453–456, Florence, Italy.

Hans P. Krings. 1986. *Was in den Köpfen von Übersetzern vorgeht*. Gunter Narr, Tübingen, Germany.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of IWSLT*, Da Nang, Vietnam.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of WMT*, pages 389–399, Copenhagen, Denmark.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of WMT*, pages 368–373, Berlin, Germany.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of NIPS*, pages 3104–3112, Montreal, Canada.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of WMT*, Brussels, Belgium.

Georges Van Slype. 1979. Critical study of methods for evaluating the quality of machine translation. Research report BR 19142 prepared for the Commission of the European Communities, Bureau Marcel van Dijk, Brussels, Belgium.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of NIPS*, pages 5998–6008, Long Beach, CA.

Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann, editors. 2017. *Proceedings of the Third Workshop on Discourse in Machine Translation*. Copenhagen, Denmark.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *Computing Research Repository*, arXiv:1609.08144.