Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation

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

Sheila Castilho **Ke Hu** **Andy Way**
ADAPT Centre
Dublin City University
Ireland
firstname.secondname@adaptcentre.ie

Abstract

We reassess a recent study (Hassan et al., 2018) that claimed that machine translation (MT) has reached human parity for the translation of news from Chinese into English, using pairwise ranking and considering three variables that were not taken into account in that previous study: the language in which the source side of the test set was originally written, the translation proficiency of the evaluators, and the provision of inter-sentential context. If we consider only original source text (i.e. not translated from another language, or translationese), then we find evidence showing that human parity has not been achieved. We compare the judgments of professional translators against those of non-experts and discover that those of the experts result in higher inter-annotator agreement and better discrimination between human and machine translations. In addition, we analyse the human translations of the test set and identify important translation issues. Finally, based on these findings, we provide a set of recommendations for future human evaluations of MT.

1 Introduction

Neural machine translation (NMT) has revolutionised the field of MT by overcoming many of the weaknesses of the previous state-of-the-art phrase-based machine translation (PBSMT) (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017). In only a few years since the first working models, this approach has led to a substantial improvement in translation quality, reported in terms of automatic metrics (Bojar et al., 2016, 2017; Sennrich et al., 2016). This has ignited higher levels of expectation, fuelled in part by hyperbolic claims from large MT developers. First we saw in Wu et al. (2016) that Google NMT was "bridging the gap between human and machine translation [quality]". This was amplified recently by the claim by Hassan et al. (2018) that Microsoft had "achieved human parity" in terms of translation quality on news translation from Chinese to English, and more recently still by SDL who claimed to have "cracked" Russian-to-English NMT with "near perfect" translation quality.[1] However, when human evaluation is used to compare NMT and SMT, the results do not always favour NMT (Castilho et al., 2017a,b).

Accompanying the claims regarding the capability of the Microsoft Chinese-to-English NMT system, Hassan et al. (2018) released their experimental data[2] which permits replicability of their experiments. In this paper, we provide a detailed examination of Microsoft's claim to have reached *human parity* for the task of translating news from Chinese (ZH) to English (EN). They provide two definitions in this regard, namely:

Definition 1. *If a bilingual human judges the quality of a candidate translation produced by a human to be equivalent to one produced by a machine, then the machine has achieved human parity.*

Definition 2. *If there is no statistically significant difference between human quality scores for a test set of candidate translations from a machine translation system and the scores for the corresponding human translations then the machine has achieved human parity.*

The remainder of the paper is organised as follows. First, we identify and discuss three potential issues in Microsoft's human evaluation, concerning (i) the language in which the source text was originally written, (ii) the competence of the human evaluators with respect to translation, and (iii) the linguistic context available to these evaluators (Section 2). We then conduct a new modified

[1] https://www.sdl.com/about/news-media/press/2018/sdl-cracks-russian-to-english-neural-machine-translation.html

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

Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 113–123
Belgium, Brussels, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64012

evaluation of their MT system on the same dataset taking these issues onboard (Section 3). In so doing, we reassess whether human parity has indeed been achieved following what we consider to be a fairer evaluation setting. We then take a closer look at the quality of Microsoft's dataset with the help of an English native speaker and a Chinese native speaker, and discover a number of problems in this regard (Section 4). Finally, we conclude the paper (Section 5) with a set of recommendations for future human evaluations, together with some remarks on the risks for the whole field of overhyping the capability of the systems we build.

2 Potential Issues

2.1 Original Language of the Source Text

The test set used by Hassan et al. (2018) (`newstest2017`) was the ZH reference from the news translation shared task at WMT 2017 (Bojar et al., 2017),[3] which contains 2,001 sentence pairs, of which half were originally written in ZH and the remaining half were originally written in EN. Figure 1 represents the WMT test set and the respective translation. The organisers of WMT 2017 manually translated each of these two subsets (files A1 and B1 in Figure 1) into the other language (B2 and A2, respectively) to produce the resulting parallel test set of 2,001 sentence pairs. Thus, Hassan et al. (2018) machine-translated 2,001 sentences from ZH into EN, but only half of them were originally written in ZH (file D1); the other half were originally written in EN, then they were translated by a human translator into ZH (as part of WMT's organisation), and this human translation was finally machine-translated by Microsoft into EN (file D2). Microsoft also human-translated the ZH reference file into EN to use as reference translations (file C - EN REF). Therefore, 50% of their EN reference comprises EN translations direct from the original Chinese (file C1), while 50% are EN translations from the human-translated file from EN into ZH (file C2), i.e. backtranslation of the original EN (A1). While their human evaluation is conducted on three different subsets (referred to as Subset-2, Subset-3, and Subset-4 in Tables 5d to 5f of their paper), since all three are randomly sampled from the whole test set, these subsets still contain around 50% of sentences originally written in ZH and around 50% originally written in EN.

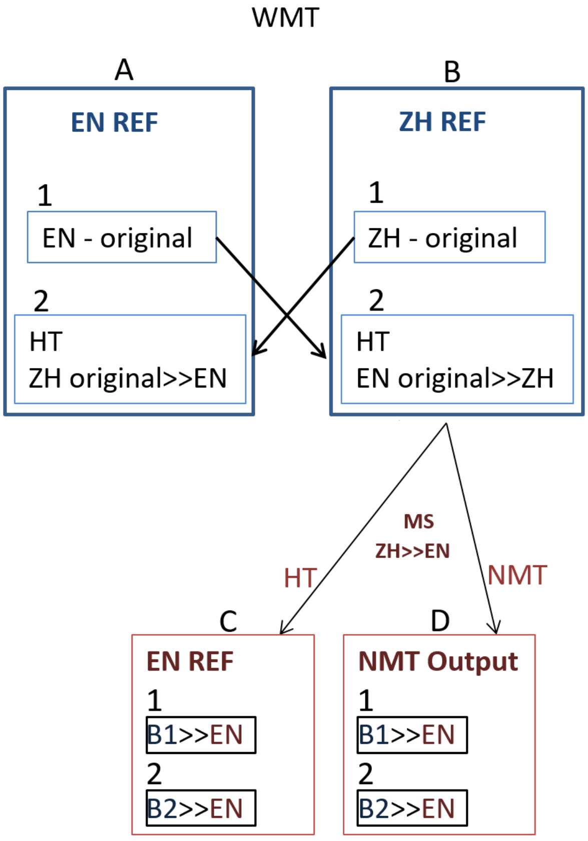


Figure 1: WMT test set and Microsoft Translation ZH-to-EN reference and MT output

We hypothesize that the sentences originally written in EN are easier to translate than those originally written in ZH, due to the simplification principle of translationese, namely that translated sentences tend to be simpler than their original counterparts (Laviosa-Braithwaite, 1998). Two additional universal principles of translation, explicitation and normalisation, would also indicate that a ZH text originally written in EN would be easier to translate. Therefore, we explore whether the inclusion of source ZH sentences originally written in EN distorts the results, and unfairly favours MT.

2.2 Human Evaluators

The human evaluation described in Hassan et al. (2018) was conducted by "bilingual crowd workers". While the authors implemented a set of quality controls to "guarantee high quality results", no further details are provided on the selection of evaluators and their linguistic expertise. In addition, no inter-annotator agreement (IAA) figures were provided. We acknowledge, however, that agreement cannot be measured using the conven-

[3] http://www.statmt.org/wmt17/translation-task.html

tional Kappa coefficient, since their human evaluation uses a continuous scale (range $[0-100]$).

It has been argued that non-expert translators lack knowledge of translation and so might not notice subtle differences that make one translation better than another. This was observed in the human evaluation of the TraMOOC project[4] in which authors compared the evaluation of MT output of professional translators against crowd workers (Castilho et al., 2017c). Results showed that for all language pairs (involving 11 languages), the crowd workers tend to be more accepting of the MT output by giving higher fluency and adequacy scores and performing very little post-editing.

With that in mind, we attempt to replicate the results achieved in Hassan et al. (2018) by redoing the manual evaluation with participants with different levels of translation proficiency, namely professional translators (henceforth referred to as experts) and bilingual speakers with no formal translation qualifications (henceforth referred to as non-experts).

2.3 Context

Hassan et al. (2018) evaluated the sentences in the testset in randomised order, meaning that sentences were evaluated in isolation. However, documents such as the news stories that make up the test set contain relations that go beyond the sentence level. To translate them correctly one needs to take this inter-sentential context into account (Voigt and Jurafsky, 2012; Wang et al., 2017a). The MT system by Hassan et al. (2018) translates sentences in isolation while humans naturally consider the wider context when conducting translation.

Our hypothesis is that referential relations that go beyond the sentence-level were ignored in the evaluation as its setup considered sentences in isolation (randomised). This probably resulted in the evaluation missing some errors by the MT system that might have been caused by its lack of inter-sentential contextual knowledge. In contrast, our revised human evaluation takes inter-sentential context into account. Sentences are not randomised but evaluated in the order they appear in the documents that make up the test set. In addition, when a sentence is evaluated, the evaluator can see both the previous and the next sentence, akin to how a professional translator works in practice. In the same spirit, concurrent work by Läubli et al. (2018) contrasts the evaluation of single sentences and entire documents in the dataset by Hassan et al. (2018), and shows a stronger preference for human translation over MT when evaluating documents as compared to isolated sentences.

3 Evaluation

3.1 Experimental Setup

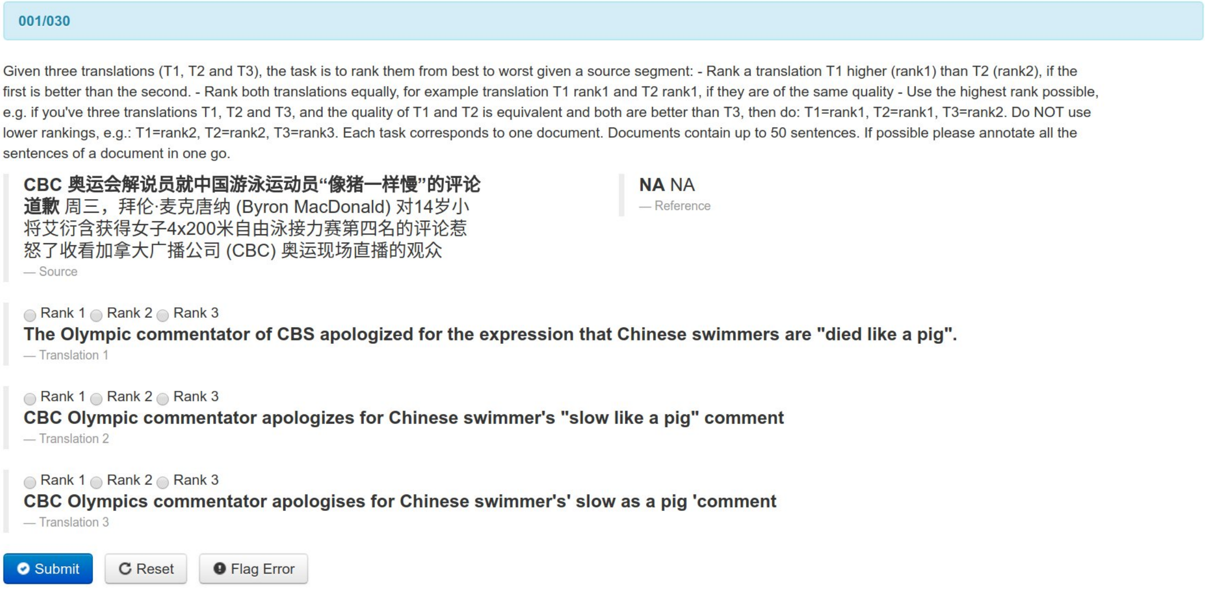
We conduct a human evaluation in which at the same time evaluators are shown a source ZH sentence and three EN translations thereof: (i) the human translation produced by Microsoft (file C in Figure 1: henceforth referred to as HT), (ii) the output of Microsoft's MT system (file D: henceforth MS), and the output of a production system, Google Translate (henceforth GG).[5] We take these three translations from the data provided by Hassan et al. (2018).

Instead of giving evaluators randomly selected sentences, they see them in order. We randomised the documents in the test set (169) and prepared one evaluation task per document, for the first 49 documents (503 sentences). Of these 49 documents, 41 were originally written in ZH (amounting to 299 sentences, with each document containing 7.3 sentences on average) and the remaining 8 were originally written in EN (204 sentences, average of 25.5 sentences per document). Evaluators were asked to annotate all the sentences of each document in one go, so that they can take inter-sentential context into account.

Rather than direct assessment (DA) (Graham et al., 2015), as in Hassan et al. (2018), we conduct a relative ranking evaluation. While DA has some advantages over ranking and has replaced the latter at the WMT shared task since 2017 (Bojar et al., 2017), ranking is more appropriate for our evaluation due to the fact that we evaluate sentences in consecutive order (rather than randomly). This can be accommodated in ranking as we can show all three translations for each source sentence together with the previous and next source sentences

[4] http://tramooc.eu/

[5] We note that in the study by Hassan et al. (2018), 9 different translations were compared: 3 reference translations, and the output from six MT systems, 4 of which were Microsoft systems (including one online), plus Google Translate and the Sogou system (Wang et al., 2017b), the best-performing system at WMT-2017. This, together with the fact that we use different methods, may affect the comparability of the results obtained to some degree.



001/030

Given three translations (T1, T2 and T3), the task is to rank them from best to worst given a source segment: - Rank a translation T1 higher (rank1) than T2 (rank2), if the first is better than the second. - Rank both translations equally, for example translation T1 rank1 and T2 rank1, if they are of the same quality - Use the highest rank possible, e.g. if you've three translations T1, T2 and T3, and the quality of T1 and T2 is equivalent and both are better than T3, then do: T1=rank1, T2=rank1, T3=rank2. Do NOT use lower rankings, e.g.: T1=rank2, T2=rank2, T3=rank3. Each task corresponds to one document. Documents contain up to 50 sentences. If possible please annotate all the sentences of a document in one go.

CBC 奥运会解说员就中国游泳运动员“像猪一样慢”的评论道歉 周三，拜伦·麦克唐纳 (Byron MacDonald) 对14岁小将艾衍含获得女子4x200米自由泳接力赛第四名的评论惹怒了收看加拿大广播公司 (CBC) 奥运现场直播的观众
— Source

NA NA
— Reference

○ Rank 1 ○ Rank 2 ○ Rank 3
The Olympic commentator of CBS apologized for the expression that Chinese swimmers are "died like a pig".
— Translation 1

○ Rank 1 ○ Rank 2 ○ Rank 3
CBC Olympic commentator apologizes for Chinese swimmer's "slow like a pig" comment
— Translation 2

○ Rank 1 ○ Rank 2 ○ Rank 3
CBC Olympics commentator apologises for Chinese swimmer's' slow as a pig 'comment
— Translation 3

Figure 2: Snapshot from the human evaluation showing the first sentence from the first document, which contains 30 sentences.

at the same time. In contrast, in DA only one translation is shown at a time, which is of course evaluated in isolation. An important advantage of DA is that the number of annotations required grows linearly (rather than exponentially with ranking) with the number of translations to be evaluated; this is relevant for WMT's shared task as there may be many MT systems to be evaluated, but not for our research as we have only three translations (HT, MS and GG). In any case, both approaches have been found to lead to very similar outcomes as their results correlate very strongly ($R \geq 0.92$ in Bojar et al. (2016)).

Our human evaluation is performed with the Appraise tool (Federmann, 2012).[6] Figure 2 shows a snapshot of the evaluation. Subsequently, we derive an overall score for each translation (HT, MS and GG) based on the rankings. To this end we use the TrueSkill method adapted to MT evaluation (Sakaguchi et al., 2014) following its usage at WMT15,[7] i.e. we run 1,000 iterations of the rankings recorded with Appraise followed by clustering (significance level $\alpha = 0.05$).

Five evaluators took part in our evaluation: two professional Chinese-to-English translators and three non-experts. Of the two professional translators, one is a native English speaker with a fluent level of Chinese, and the other is a Chinese native speaker with a fluent level of English. The three non-expert bilingual participants are Chinese native speakers with an advanced level of English. These bilingual participants are researchers in NLP, and so their profile is similar to some of the human evaluators of WMT, namely MT researchers.[8]

All evaluators completed all 49 documents, except the third non-expert, who completed the first 18. Similarly, all evaluators ranked all the sentences in the documents they evaluated, except the second professional translator, who skipped 3 sentences. In total we collected 6,675 pairwise judgements.

3.2 Results

3.2.1 Original Language

To find out whether the language in which the source sentence was originally written has any effect on the evaluation, we show the resulting Trueskill scores for each translation taking into account all the sentences in our test set versus considering the sentences in two groups according to the original language (ZH and EN). The results are shown in Table 1.

Regardless of the original language, GG is the lowest-ranked translation, thus providing an indi-

[6] https://github.com/cfedermann/Appraise

[7] https://github.com/mjpost/wmt15

[8] It is an open question as to whether using bilingual NLP researchers may affect the results obtained. While we follow the practice of WMT here – which differs from the approach taken by Hassan et al. (2018), who used bilingual crowd workers – we intend in future work to investigate this further.

Rank	Original language		
	Both $n = 6675$	ZH $n = 3873$	EN $n = 2802$
1	HT 1.587*	HT 1.939*	MS 1.059
2	MS 1.231*	MS 1.199*	HT 0.772*
3	GG -2.819	GG -3.144	GG -1.832

Table 1: Ranks of the translations given the original language of the source side of the test set shown with their Trueskill score (the higher the better). An asterisk next to a translation indicates that this translation is significantly better than the one in the next rank.

cation that the quality obtainable from the MS system is a notable improvement over state-of-the-art NMT systems used in production. We observe that HT outperforms significantly MS when the original language is ZH, but the difference between the two is not significant when the original language is EN. Hence, we confirm our hypothesis that the use of translationese as the source language distorts the results in favour of MS.

Next, we check whether this effect of translationese is also present in the evaluation by Hassan et al. (2018). To this end, we concatenate all their judgments and model them with mixed-effects regression. Our dependent variable is the score, scaled down from the original range $[0, 100]$ to $[0, 1]$, which we aim to predict with one continuous predictor – sentence length – and two factorial independent variables: translation (levels HT, MS and GG) and original language (levels ZH and EN). The identifiers of the sentence and the annotator are included as random effects. We plot the interaction between the translation and the original language of the resulting model in Figure 3. HT outperforms MS by around 0.05 absolute points for sentences whose original language is ZH. However this gap disappears for source sentences originally written in EN, where we see that the score for MS is actually slightly higher than that of HT, though the difference is not significant. We observe a clear effect of translationese (EN): compared to ZH, the scores of both MT systems increase substantially (GG over 10% absolute and MS over 6% absolute), while the HT score increases only very slightly.

Our hypothesis was theoretically supported by the simplification principle of translationese. Applied to the test data, this would mean that the portion originally written in ZH is more complex than the part originally written in EN. To check whether this is the case, we compare the two subsets of the test set using a measure of text complexity, type-token ratio (TTR). While both subsets contain a similar number of sentences (1,001 and 1,000), the ZH subset contains more tokens (26,468) than its EN counterpart (22,279). We thus take a subset of the ZH (840 sentences) containing a similar amount of words to the EN data (22,271 words). We then calculate the TTR for these two subsets using bootstrap resampling. The TTR for ZH ($M = 0.1927$, $SD = 0.0026$, 95% confidence interval $[0.1925, 0.1928]$) is 13% higher than that for EN ($M = 0.1710$, $SD = 0.0025$, 95% confidence interval $[0.1708, 0.1711]$).

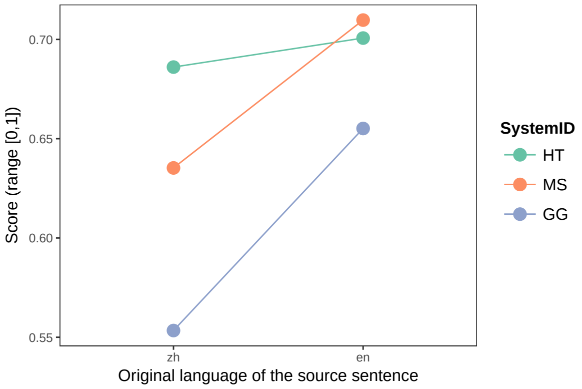


Figure 3: Interaction between the MT system (levels HT, MS and GG) and the original language of the source sentence (levels ZH and EN).

Given the findings of this experiment, in the remainder of the paper we use only the subset of the test set whose original language is ZH.

3.2.2 Evaluators

To find out whether the translation expertise of the evaluator has any effect on the evaluation, we show the resulting Trueskill scores for each translation resulting from the evaluations by non-expert versus expert translators. The results are shown in Table 2. The gap between HT and MS is considerably wider for experts (2.2 vs 1.2) than for non-experts (1.3 vs 0.9). We link this to our expectation, based on the previous finding by Castilho et al. (2017c), that non-experts are more lenient regarding MT errors. In other words, non-experts disregard translation subtleties in their evaluation, which leads to the gap between different translations – in this case HT and MS – being smaller. In Section 4 we explore this further by means of a qualitative analysis.

Rank	Translators		
	All $n = 3873$	**Experts** $n = 1785$	**Non-experts** $n = 2088$
1	HT 1.939*	HT 2.247*	HT 1.324
2	MS 1.199*	MS 1.197*	MS 0.94*
3	GG -3.144	GG -3.461	GG -2.268

Table 2: Ranks and Trueskill scores (the higher the better) of the three translations for evaluations carried out by expert versus non-expert translators. An asterisk next to a translation indicates that this translation is significantly better than the one in the next rank.

Trueskill provides not only an overall score for each translation but also its confidence interval. We expect these to be wider for the annotations by non-experts than those annotations given by experts, which would indicate that there is more uncertainty in the rankings by non-experts. Figure 4 shows the scores for each translation by experts and non-experts, i.e. the same values that were shown in Table 2, now enriched with their 95% confidence intervals.

The sum of the confidence scores for the three translations is just 0.33% higher for non-experts (3.076) than for experts (3.066). However, it is worth mentioning that, compared to the width of the intervals for experts, those for non-experts are considerably wider for HT (16% relative difference) while they are similar or smaller for MT (1% and -11% relative differences for GG and MS, respectively).

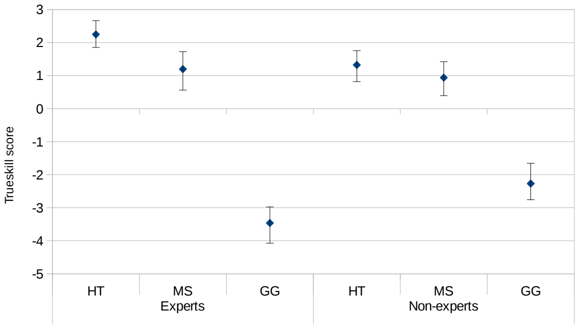


Figure 4: Trueskill scores of the three translations by experts and non-experts together with their confidence intervals.

We now look at inter-annotator agreement (IAA) between experts and non-experts. We compute the Kappa (κ) coefficient (Cohen, 1960), as done at WMT 2016 (Bojar et al., 2016, Section 3.3):[9]

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ represents the proportion of times that the annotators agree, and $P(E)$ the proportion of times that the annotators are expected to agree by chance.

As expected, the IAA between professional translators ($\kappa = 0.254$) is notably higher, 95% relative, than that between non-experts ($\kappa = 0.130$).[10] As we have three non-experts, we can calculate the IAA not only among the three of them but also between all three pairs of non-expert annotators; all of the resulting coefficients (0.057, 0.135 and 0.195) are lower than that between experts (0.254).

To the best of our knowledge, this is the first time that IAA of professional translators and non-experts has been compared for the human evaluation of MT. In related work, Callison-Burch (2009) compared the agreement level of two types of non-expert translators: MT developers (referred to in that paper as 'experts') and crowd workers. He showed that crowd workers can reach the agreement level of MT researchers using multiple workers and weighting their judments. That said, both types of non-experts conducted human evaluations for WMT13 (Bojar et al., 2013) and the IAA rates of the crowd were well below those of the researchers.

4 Analyses

As mentioned previously, we have examined the quality of the test sets, both originally written in ZH and originally written in EN and their respective translations. An English native speaker analysed both the original EN version from the WMT set (file A1 in Figure 1) and the human translation of the set originally written in ZH performed by Microsoft (file C2). A Chinese native speaker, who is fluent in English and has experience with translation from EN into ZH, analysed the original

[9] https://github.com/cfedermann/wmt16/blob/master/scripts/compute_agreement_scores.py

[10] Due to the fact that one non-expert evaluated only 18 out of the 49 documents, the IAA calculations consider only the first 18 documents. If we consider all 49 documents, the trend remains the same; the IAA for the two experts is higher than that for the two non-experts who evaluated all the documents: 0.265 vs 0.196.

ZH versions (file B1) as well as the human translation of the set originally written in EN performed by the WMT organisers (file B2).

4.1 Original English

Regarding the English original (file A1 in Figure 1), the analysis showed that apart from a few grammar errors, the test set appeared to be fluent and grammatical. Examples of grammatical errors in the original EN files are:

i) "The idiot didn't realize they were still on the air"
ii) "Soon after, Scott Russel who was hosting CBC's broadcast apologized on-air for MacDonald's comment, saying: 'We apologize the comment on a swim performance made it to air.' "

In example i) "on air" should be used instead of "on the air", while in the example ii) a missing "that" should be used after "apologize". Nonetheless, these errors did not affect the ZH translation (file B2) or the following backtranslation (C2) into EN. Our hypothesis is that because the test set is news content, it also contains tweets (such as example i)) and quotes from speech interviews (such as example ii)), which are more likely to contain grammatical errors.

4.2 Chinese Translation

Regarding the human translation into ZH performed by WMT (file B2 in Figure 1), most of the sentences contained grammatical errors and/or mistranslations of proper nouns. Furthermore, although some translations were grammatically correct and accurate, they were not fluent. When the ZH-translated sentences were compared against the source (A1), the translations were mostly accurate. However, when analyzed on their own without the source, they sound disfluent:

iii)
EN original (A1): A front-row seat to the stunning architecture of the Los Angeles Central Library
ZH (B2):洛杉矶中央图书馆的惊艳结构先睹为快

iv)
EN original (A1): An open, industrial loft in DTLA gets a cozy makeover
ZH (B2): DTLA的开放式工厂阁楼进行了一次舒适的改造。

In example iii), although the ZH translation has fully transferred the meaning of the source text, it contains word-order errors which makes the translation disfluent since the verb phrase "先睹为快" (take a look) is placed after the object (library). One possible translation for that is "抢先目睹洛杉矶中央图书馆的惊艳结构" because the ZH language syntax requires the verb to be placed before the object.

In example iv), the ZH translation contains a grammatical error in the word "进行", which would imply that the loft is carrying out a makeover. In addition, the adjective "舒适的" (cosy) cannot be used to describe "改造" (makeover). One possible translation for the English sentence is "DTLA的开放式工业阁楼被改造的很舒适".

Given this analysis, we speculate that the translation of the EN original files into ZH might not have been performed by an experienced translator, but rather exemplify either human translation performed by an inexperienced translator, or poorly post-edited MT.

4.3 English Translation

Regarding the EN reference files translated by Microsoft (file C2 in Figure 1), many of the sentences contained grammatical errors (such as word order, verb tense and missing prepositions) as well as mistranslations.

v)
EN original (A1): A front-row seat to the stunning architecture of the Los Angeles Central Library
ZH (B2):洛杉矶中央图书馆的惊艳结构先睹为快
EN (C2): Take a look of the astounding architecture of the Los Angeles Central Library.

GG: The stunning structure of the Los Angeles Central Library
MS: A sneak peek at the stunning architecture of the Los Angeles Central Library

vi)
EN original (A1): An open, industrial loft in DTLA gets a cozy makeover
ZH (B2):DTLA的开放式工厂阁楼进行了一次舒适的改造。
EN (C2): A comfortable makeover was provided

to the open factory building design of DTLA.

GG: DTLA’s Open factory loft has a comfortable makeover.
MS: DTLA’s open-plan factory loft has undergone a comfortable makeover.

In example v), the EN translation of the ZH source[11] analyzed previously is translated with the wrong preposition, i.e. ‘look of’ instead of ‘look at’. None of the professional translators considered the reference worse than the MS output; while one translator and one non-expert considered it ‘as good’ as the MS output, the other considered it better than MS but worse than GG. Regarding the non-expert assessment, two of them considered the HT to be as good as MS and better than GG, and one considered the HT to be worse than MS but better than GG.

In example vi), the EN translation (C2) of the ZH source (B2) does not have all the information expressed in ZH as the word ‘loft’ (阁楼) is not translated. Moreover, the EN translation refers to an architectural design makeover of the building rather than an interior makeover of an attic. Both professional translators considered the EN reference to be worse than the MS output. As far as the non-experts are concerned, two of them considered the HT to be worse than MS, while one considered it to be ‘as good’. This provides qualitative evidence that non-experts may be more tolerant of translation errors than professional translators.

Another example of such behaviour is the following:

vii)
EN original (A1): Learn more about the history of downtown’s Central Library as the Society of Architectural Historians/Southern California Chapter hosts a salon with Arnold Schwartzman and Stephen Gee, authors of the new book ”Los Angeles Central Library: A History of its Art and Architecture
ZH (B2): 美国建筑史学家学会南加利福尼亚洲分会与新书《洛杉矶中央博物馆：其艺术与建筑历史》的作者阿诺·斯瓦茨曼和史蒂芬·吉举办了一场沙龙。观众们可通过此次活动进一步了解市中心中央图书馆的历史
EN (C2):A salon will be hosted by Southern California Branch of Society of Architectural Historians and the co-authors of Los Angles Central Museum: Art and Architectural History, Arnold Schwarzman and Stephen Gee. It will deliver more knowledge of the Central Library to the participants

GG: The Southern California branch of the American Institute of Architectural Historians has held a salon with 阿诺·斯瓦茨曼 and 史蒂芬·吉, author of the Los Angeles Central Museum: its art and architectural history. Through this event, viewers can learn more about the history of Central Library in the city centre
MS: The Southern California chapter of the American Society of Architectural Historians and the authors of a new book, ”Los Angeles Central Museum: Its Art and Architectural History,” Arnold Schwartzman and Steven Gee, hosted a salon at which viewers learned more about the history of the Central Library in the city center

In example vii), regarding the ZH source (B2), in addition to having the first sentence translated into past tense – whereas the EN original (A1) shows the salon event is happening in the near future – it also contains a typo ‘洲’ which means ‘continent’ instead of ‘state’ ‘州’. Even though the typo does not affect the EN translation (C2), it shows that the quality of the ZH translation is not as high as would be expected of professional human translators. Regarding the EN translation (C2), while the first sentence is mostly fluent – although it contains a typo in ‘Angles’ (Angeles) and lacks the article ‘the’ before the proper noun in the first sentence – the second sentence lacks fluency and contains errors of omissions and mistranslations. For example, the words “downtown” and “history” (市中心 and 历史, respectively) were not transferred over to the EN reference (C2). Furthermore, the word ‘viewers’ in the ZH translation (观众们) was mistranslated as ‘participants’. Nonetheless, the EN translation (C2) is able to capture the correct tense of the sentence since the second sentence in the ZH translation (B2) is ambiguous regarding verbal tense. The MS translation does a better job in keeping the fluency throughout the sentence even though it mis-

[11] It is important to note that the translators did not have access to the original EN (A1) and so the ZH file (B2) was used as the source.

translates the tense of the source in the past tense. Both professional translators assessed the HT as worse than MS, whereas two of the non-experts considered it to be as good as MS and better than GG. The third non-expert considered the HT to be worse than both MT systems. This example shows that the level of expertise of the evaluators may have an effect on the evaluation given that non-experts are wrongly more tolerant of MT errors.

Similarly to the ZH translation (B2) of the English original, we speculate that the EN translation (C2) of the ZH files is more likely a human translation performed by an inexperienced translator, or even a poorly post-edited machine translation; even if the translation was performed by an experienced translator, such that the ZH source (B2) contained errors or was disfluent, a professional translator would surely be more meticulous and fix such errors before rubber-stamping the translations.

5 Conclusions and Future Work

This paper has reassessed a recent study that claimed that MT has reached human parity for the translation of news from Chinese into English, considering three variables that were not taken into account in that previous study: (i) the language in which the source side of the test set was originally written, (ii) the translation proficiency of the evaluators, and (iii) the provision of inter-sentential context.

The main findings of this paper are the following:

- If we consider the subset of the test set whose source side was originally written in ZH, there is evidence that human parity has not been achieved, i.e. the difference between the human and the machine translations is significant. This is the case both in our human evaluation and in Microsoft's.

- Having translationese (ZH translated from EN in our study) as input, compared to having original text, results in higher scores for MT systems in Microsoft's human evaluation.

- Compared to judgments by non-experts, those by professional translators have a higher IAA and a wider gap between human and machine translations.

- We have identified issues in the human translations by both WMT and Microsoft. These indicate that these translations were conducted by non-experts and that were possibly post-edited MT output.

There is little doubt that human evaluation has played a very important role in MT research and development to date. As MT systems improve – as exemplified by the progress made by Hassan et al. (2018) over state-of-the-art production systems – and thus the gap between them and human translators narrows, we believe that human evaluation, in order to remain useful, needs to be more discriminative. We suggest that a set of principles should be adhered to, partly based on our findings, which we outline here as recommendations:

- The original language in which the source side of the test sets is written should be the same as their source language. This will avoid having spurious effects because of having translationese as MT input.

- Human evaluations should be conducted by professional translators. This allows fine-grained nuances of translations to be taken into account in the evaluation and should result in higher inter-annotator agreement.

- Human evaluations should proceed taking the whole document into account rather than evaluating sentences in isolation. This allows for intersentential phenomena to be considered as part of the evaluation.

- Test sets should be translated by experienced professional translators from scratch.

We are confident that adhering to these principles is sensible under any translation conditions. Of course, if the test set is faulty, then in claiming that one's MT system outperforms one's competitors, there is a risk that what one is actually demonstrating is the contrary, as if automatic evaluation metrics demonstrate a higher score, what that could be denoting is that one's output is actually closer to the faulty test set than producing better output in terms of improved translation quality *per se*. Of course, this has consequences not just for the study in this paper, but for all shared tasks: past, present, and future.[12]

[12] Ideally, it would be great if multiple references were also

Should material be made available by Google, SDL or any other MT developers who claim ‘human parity’ or the like, we would be very happy to apply these principles in subsequent rigorous evaluations of actual demonstrable improvements in translation quality. One thing is certain; as Way (2018) observes, “those of us who have seen many paradigms come and go know that overgilding the lily does none of us any good, especially those of us who have been trying to build bridges between MT developers and the translation community for many years.” We trust that our findings in this paper demonstrate that while MT quality does seem to be improving quite dramatically, human translators will continue to find gainful employment for many years to come, despite somewhat grandiose claims to the contrary.

On a final note, we acknowledge that our conclusions and recommendations are somewhat limited in that they are derived from experiments on just one language direction and five evaluators. Therefore we plan as future work to conduct similar experiments on additional language pairs with a higher number of evaluators. In the spirit of Hassan et al. (2018), without which this paper would not have been possible, we too make publicly available our evaluation materials, the anonymised human judgments and the statistical analyses thereof.[13]

Acknowledgments

We would like to thank the five expert and non-expert translators that took part in this study. We are also grateful for valuable comments on this paper from Hany Hassan, the lead author on the Microsoft paper. This research was partially supported by the iADAATPA project funded by CEF research action (2016-EU-IA-0132) under grant agreement No.1331703. The ADAPT Centre for Digital Content Technology at Dublin City University is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/ 2106) and is co-funded under the European Regional Development Fund.

References

Luisa Bentivogli, 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.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

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 the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, 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. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon’s Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP ’09, pages 286–295, Singapore.

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(1):109–120.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota 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 *MT Summit 2017*, pages 116–131, Nagoya, Japan.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Andy Way, Panayota Georgakopoulou, Maria Gialama, Vilelmini Sosoni, and Rico Sennrich. 2017c. Crowdsourcing for NMT evaluation: Professional

available, but the point remains that if these are poor quality human translations, then this is likely to skew results still further.

[13] https://github.com/antot/human_parity_mt

translators versus the crowd. In *Translating and the Computer 39*, London. https://www.asling.org/tc39/?page_id=3223.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado.

Hany Hassan, 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. https://arxiv.org/abs/1803.05567.

Samuel Läubli, 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.

Sara Laviosa-Braithwaite. 1998. Universals of translation. In *Routledge Encyclopedia of Translation Studies*, pages 288–291. Routledge, London.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient Elicitation of Annotations for Human Evaluation of Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT16)*, pages 371–376, Berlin, Germany.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain.

Rob Voigt and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montrèal, Canada.

Longyue Wang, Zhaopeng Tu, Andy Way, and Liu Qun. 2017a. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark.

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017b. Sogou neural machine translation systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, WMT 2017*, page 410–415, Copenhagen, Denmark.

Andy Way. 2018. Machine translation: Where are we at today? In Angelone E, Massey G, and Ehrensberger-Dow M, editors, *The Bloomsbury Companion to Language Industry Studies*. Bloomsbury, London. In press.

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. https://arxiv.org/abs/1609.08144.

Appendix: Evaluator Instructions

Given three translations (T1, T2 and T3), the task is to rank them from best to worst given a source segment:

- Rank a translation T1 higher (rank1) than T2 (rank2), if the first is better than the second.
- Rank both translations equally, for example translation T1 rank1 and T2 rank1, if they are of the same quality.
- Use the highest rank possible, e.g. if you've three translations T1, T2 and T3, and the quality of T1 and T2 is equivalent and both are better than T3, then do: T1=rank1, T2=rank1, T3=rank2. Do NOT use lower rankings, e.g.: T1=rank2, T2=rank2, T3=rank3.

Each task corresponds to one document. Documents contain up to 50 sentences. If possible please annotate all the sentences of a document in one go.