The Suboptimal WMT Test Sets and Their Impact on Human Parity

Ahrii Kim[1†], **Yunju Bak**[2†], **Jimin Sun**[3§*], **Sungwon Lyu**[4†], **Changmin Lee**[5†]

[†]Kakao Enterprise [§]Carnegie Mellon University
{ria.i, juliette.y, james.ryu, louis.cm}@kakaoenterprise.com
jimins2@cs.cmu.edu

Abstract

With the advent of Neural Machine Translation, the more the achievement of human-machine parity is claimed at WMT, the more we come to ask ourselves if their evaluation environment can be trusted. In this paper, we argue that the low quality of the source test set of the news track at WMT may lead to an overrated human parity claim. First of all, we report nine types of so-called *technical contaminants* in the data set, originated from an absence of meticulous inspection after web-crawling. Our empirical findings show that when they are corrected, about 5% of the segments that have previously achieved a human parity claim turn out to be statistically invalid. Such a tendency gets evident when the contaminated sentences are solely concerned. To the best of our knowledge, it is the first attempt to question the "source" side of the test set as a potential cause of the overclaim of human parity. We cast evidence for such phenomenon that according to sentence-level TER scores, those trivial errors change a good part of system translations. We conclude that to overlook it would be a mistake, especially when it comes to an NMT evaluation.

Keywords: Human Parity, NMT Evaluation, WMT

1. Introduction

Since the astonishing performance of neural machine translation (NMT) (Bahdanau et al., 2014; Cho et al., 2014; Vaswani et al., 2017) has caused a sensation in the MT research community, winning back the state-of-the-art position from Statistical Machine Translation (Bentivogli et al., 2016), it did not take more than two years for the newcomer to reach a human-level translation quality in public. As a front runner, a group of researchers at Microsoft claimed that their models were at human parity in a Chinese→English test set distributed by Conference on Machine Translation (WMT) 2017 (Hassan et al., 2018). In detail, their three NMT models were evaluated via a source-based direct assessment against three levels of human translations (HT) and were positioned at the top rank where the HT of the highest quality was situated. Subsequently, Bojar et al. (2018) claimed another human parity in the English→Czech translation at WMT 18, and Barrault et al. (2019) even declared superhuman performance in the German→English and English→Russian translation at WMT 19.

The fact that such results were acquired from the best practices of WMT evaluation protocol should make these claims more trustworthy, considering the authority of the campaign mentioned above as one of the most active gatherings in terms of MT evaluation since 2006. Despite their glory, however, these claims were regarded as a "hyperbolic" move and received with skepticism by many in the relevant field, who at once revisited some of the most representative exercises (Läubli et al., 2018; Toral et al., 2018; Graham et al., 2019; Toral, 2020; Graham et al., 2020; Läubli et al., 2020). The gist of their work was that the current evaluation standards of WMT were insufficient to fairly assess the high performance of NMT models for the following motives:

- Sentence-level assessment without context information
- Assessment by an inexpert group
- Reference translation of low quality
- Adverse effect of translationese

They proved that when the evaluation surroundings were improved, the gap between HT and MT was noticeably widened and, therefore, such human parity claims were premature.

In a like manner, we cast doubt on the human parity claims by suggesting another erroneous evaluation environment of WMT: **contamination of the source text of the test sets**. Our primary interest lies in the influence of the polluted test set on assessing the human parity level. We, therefore, formulate a hypothesis as such:

> *Albeit minor, contaminants in a source text of a test set will bring out a false human parity claim.*

To this aim, characteristics of the contaminants are identified from two perspectives: technical (Section 3.2) and topical (Section 3.3). We manually detect 101 technical contaminants and raise a question about topical diversity by evaluating the theme of the test samples with a BERT-based news topic classifier. While a perfect scenario of "a clean test set" would be to remove both contaminants, we confine ourselves to the

[*]Work done during the author's internship at Kakao Enterprise.

	Document	Sentence	Distinct Word	Paragraph	Sentence/Paragraph	
					Mean	Max
English I	63	1,000	4,970	-	-	-
English II	61	1,002	5,040	-	-	-
English III	130	2,048	7,892	1,418	1.44	9

Table 1: Statistics of English source texts of WMT 20 test set. The paragraph-related information is only given in English III.

technical contaminants in this study. To verify their adverse effect on an NMT evaluation, we conduct a pairwise relative ranking (RR) human evaluation comparing two online NMT models and reference translation on the English→Korean translation with 15 researchers (Section 4). We report that some of the human parity claims become invalidated when the contaminants are eliminated.[1]

2. Related Works

It is not a new discovery that the test sets released by WMT contain errors, but it is unusual to expect that they will have an adverse effect on the evaluation. The most relevant work is found concerning a translationese issue in the test set of WMT 18. The term *translationese* refers to a particular feature found in a translated text. As Graham et al. (2020) summarize, these features are mainly characterized as simplified use of language and perfect grammaticality. Toral et al. (2018) and Läubli et al. (2020) demonstrate that the construction pipeline of WMT test sets posed the inherent dangers of containing translationese, which led to a more favorable judgment for MT. They conducted a RR evaluation comparing two MT models and one HT on three type of texts — i) a complete set, ii) source-language-sourced set (Chinese), and iii) target-language-sourced (English) set— provided by Hassan et al. (2018) of Microsoft. Results proved that when translationese was removed from the test set (type iii), Microsoft's NMT model that had previously achieved human parity was significantly below the human level. Since then, many reconfirmed its critical influence on NMT evaluations (Graham et al., 2019; Graham et al., 2020; Edunov et al., 2019). WMT also acknowledged that translationese should be omitted from all test sets, and it has been avoided since WMT 2019.

As a matter of fact, the quality issue of their dataset had been discussed indeed in terms of human reference translation of WMT. Hassan et al. (2018) admitted that they had to reconstruct a new reference translation of higher quality for their experiment, because WMT-provided HT created from crowdsourcing seemed to be of inadequate quality. Toral et al. (2018) criticized the quality of WMT's Chinese-English translations, informing on their grammatical as well as syntactic errors. They commented that the reference must have been created by "in-experienced" translators or by post-editing. Läubli et al. (2020) ran a pairwise RR evaluation on Adequacy and Fluency apart, in a binary way: (MT, HT_2) and (HT_1, HT_2). HT_1 was obtained from Hassan et al. (2018), known to be of high quality, and HT_2 was provided by a vendor. Results showed that HT_1 was significantly preferred to HT_2. Therefore, although MT achieved human parity over HT_2, such a claim could be on a false basis. In this respect, Toral et al. (2018) stressed the importance of a discriminative test set in an evaluation for robust models like NMT.

We deviate from those studies in that our focal point is the original source text of a test set. From our understanding, it is the first study to raise a quality issue of the source side of the test sets of WMT. While consulting the experimental setups of the aforementioned studies, we demonstrate that the contaminants in the original text harm the whole environment, inducing an overclaim of the human-machine parity. We strongly argue that such an issue should not be taken for granted when it comes to an NMT evaluation.

3. Contaminants

3.1. WMT 20 Test Set

In the WMT's news track, source texts of the test set were prepared in eight languages, inclusive of English. In 2020, three types of English source texts were constructed, each of which served as a basis for a test set of different language pairs (Barrault et al., 2020). While our experiment employs English III, we offer a holistic view of all English source texts at WMT in Table 1.

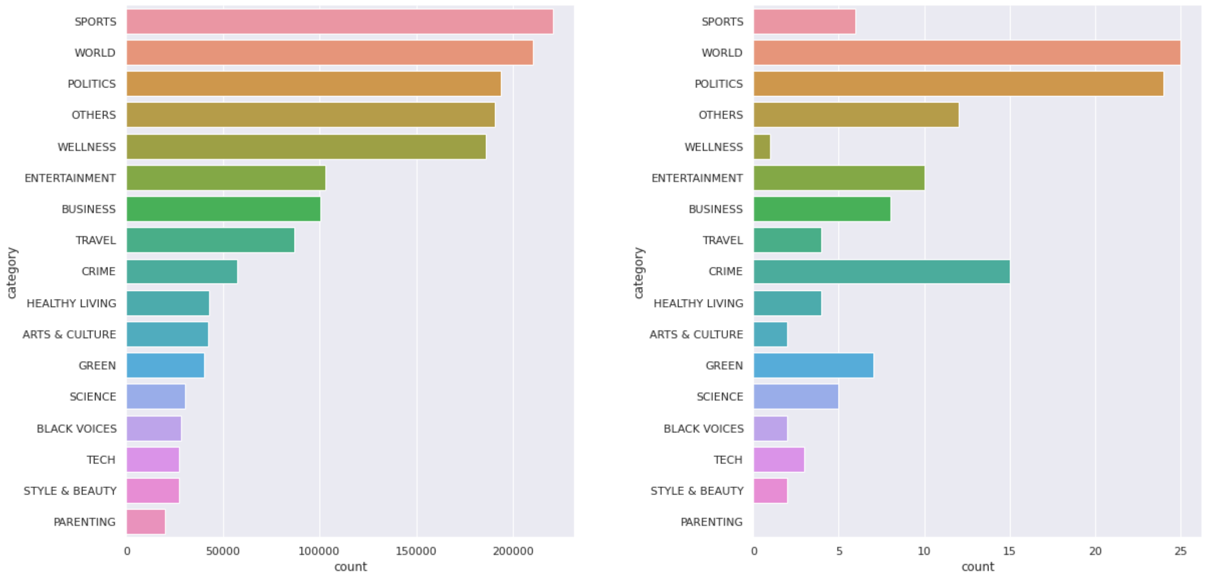
Technically speaking, the number of sentences with paragraph-split is counted with the Moses sentence parser (Koehn et al., 2007). The tokenized texts are lowercased with the Moses tokenizer and the number of distinct words was counted. As Table 1 displays, English III covers a broader range of data with more varied vocabularies and a larger scale.

3.2. Technical Contaminants

We detect *contaminants*, as we call them, on a surface level. We report that out of 130 documents, 42 documents are contaminated either slightly or considerably. Half of the cases are a single presence ($n = 21$), while the rest are multiple appearances ($n = 21$) in one document when examined on a sentence basis.

It is speculated that the majority of the technical contaminants is originated from an incorrect web-crawling, or more explicitly, the absence of meticulous

[1] Link to our code is available at `https://github.com/ahrii-kim/suboptimal_test_set`



(a) WMT20 English news-crawl

(b) WMT20 test English III

Figure 1: Topic classification of (a) training set and (b) test set of WMT 20.

Error Category	**Count**
Quotation marks	64
Apostrophe	14
Omission	50
Cultural difference*	30
Spacing	9
Typo	9
Omission of period	8
Missing headlines	4
Caption	3
Irrelevant content	3
Grammar	1
Out-of-context extraction*	18
Total	**101**

Table 2: Type of contaminants in the English III test set of WMT 20. Asterisks(*) are excluded from the total counts.

posterior inspection of the web-crawled texts. As in Table 2, we have spotted a total of 101 contaminants in the data set and classified them into nine types. It is noticeable that about 63.4% of the contaminants stem from a misuse of quotation marks. Some of the most intriguing categories are discussed in this section.

Quotation Marks It accumulates 63.4% of the total contaminants. In detail, the end quotes are mostly omitted (78%). 22% are incorrectly used as an apostrophe. In addition, we have revised 15 pairs of single quotes ($n = 30$) to double quotes as in the American manner (Category: Cultural difference) in an effort to see their impact on the result, but we notice that such case is not considered as a contaminant in the statistics.

Spacing In most cases, two words are mistakenly attached without spacing.

Typo It includes simple typos such as: *monitorigin* → *monitoring*, *he* → *He*, *n* → *in*, *Laszlo Trocsanyi* → *László Trócsányi*. Other cases seem more influential to the content when translating: *though* → *through*, *other* → *another*, *the* → *they*.

Missing Headlines Unlike other documents that always start with a headline, four documents do not have it. In order to provide an equal evaluation environment, we take this case as a contaminant. By appending their headlines through a manual web search, we assure a balance in our after-test-set.

Caption Either captions for photos or a leaflet are inserted in the middle of the documents. As they interrupt the context, they are considered as a contaminant in this paper.

Irrelevance It refers to an irrelevant segment to the context, such as a journal's date, name or city. Not only is their existence useless in an evaluation, but some of them are separated as one segment by mistake, to make matters worse.

Out-of-Context Extraction Unlike other documents extracted from the beginning of an original article, two cases in the test set start from the middle. It is problematic for humans in that it hinders the translators' understanding of the context. While it is a contaminant, we omit it from the final count.

3.3. Topical Contaminants

Together with the technical contaminants, we argue that the test set is topically imbalanced. We observe that the documents repeatedly cover similar events or deal with a limited group of figures. We denominate such phenomena as *topical contaminants* and confirm their existence with the help of a Topic Classification algorithm. Under the assumption that the aim of the news track at WMT is to evaluate MT models on a diversified testbed, we point out that a good test set should represent the training set distribution proportionally. Their statistical impact on NMT evaluation, however, is out of the scope of this study and should be further verified in a more detailed setup.

We train a classifier by fine-tuning BERT (Devlin et al., 2018) on a News Category Data Set[2] from Kaggle that consists of around 200,000 news headlines and short descriptions collected from HuffPost during 2012-2018. The original data set contains 41 topic categories, but we have merged some of the overlapped categories such as "*worldpost*" and "*the worldpost*" or "*art*" and "*art & culture*" under the category of "Representatives." The bottom 20 categories are also combined to the category of "Others." On such criteria, we compare the WMT 20 English news-crawl training set containing 1,600,000 documents and the given test set. When the distribution of the topics in both texts is compared, the result shows that the test set (b) is markedly disproportional to the distribution of the training set (a) of Figure 1. Moreover, it turns out that the given test set is considerably biased toward a few categories, including "*world*," "*politics*" or "*crime*." The least popular topic, on the other hand, would be "*parenting*," "*wellness*," or "*art & culture*."

4. Evaluation Setup

We hypothesize that contaminants of an original source text make the human parity claim more reachable, especially in terms of NMT models. We test the given hypothesis by conducting RR on two different test sets: an originally-WMT-distributed test set and its revised version. They are named after "Before Test Set (BTS)" and "After Test Set (ATS)," respectively. We believe that this type of data set construction allows us to demonstrate that the contaminants are the sole factor of the outcome.

Language Pair The translation direction is English → Korean.

Data set We use a source text of English-III-typed WMT 20 test set built for German, Czech, and Chinese (Barrault et al., 2020). The 21 documents (437 sentences) with two or more contaminants described in Section 3.2 are selected and provide a foundation for constructing BTS and ATS, totaling 874 sentences.

System translations of the two NMT models are created on July 21, 2021, all on the same day to avoid a possible temporal influence on the result. For the experiment, a mixture of BTS and ATS are shuffled document-wise to create HITs of around 100 segments each, a portion for one annotator, containing both versions as evenly as possible. Such a mixture is intended for every annotator to judge both of the data set in the assessment. In total, 10 HITs are prepared, and each HIT is assigned to two annotators. Evaluation data collected from this experiment are displayed in Table 3. In summary, we have collected 5,244 pairwise judgments.

Pair	Sys	Assess	Assess/Sys	Assess/Sent
en→ko	3	5,244	1,748	2

Table 3: Amount of collected assessment data per system (sys) and sentences (sent).

NMT Models We employ two online NMT models, anonymously denominated as MT$_Y$ and MT$_Z$. We make a special effort to select the best-performing engines for this language pair via a preliminary test, as this evaluation focuses on human parity.[3] Their translations are contrasted to a Korean reference translation created initially by a vendor and manually revised afterward by one of our researchers who has been working as a professional translator. We have endeavored to build a reference of the highest quality.

RR For each source sentence, annotators are asked to perform a pairwise ranking of three candidate translations —HT and two system translations (MT$_Y$, MT$_Z$)— from best ($1 = best$) to worst ($3 = worst$), with ties allowed. A sentence is displayed in the order of articles along with one previous/following sentence. The task is prepared on TAUS DQF [4].

Computation The result is extracted from the two-sided Sign Tests suggested by Läubli et al. (2018) or Toral (2020), where MT is better, HT is better, and their ties are calculated. Their method is designed for a binary comparison of two candidates; we iterate two independent rounds, such that (MT$_Y$, HT) and (MT$_Z$, HT) can be contrasted on the same scenario. At the end, the statistical significance is verified by a two-sided binomial test. More details in terms of the given methodology are referred to Läubli et al. (2018).

We also take a look at the trend of an absolute score of each model as guided by TAUS[5]. The score is computed by weighing each rank with different points and normalizing them by the total number of rankings, as in Equation 1. The best absolute result, therefore, is $score = 3$.

$$Score = \frac{(1_{st} * 3p) + (2_{nd} * 2p) + (3_{rd} * 1p)}{\#of rankings} \quad (1)$$

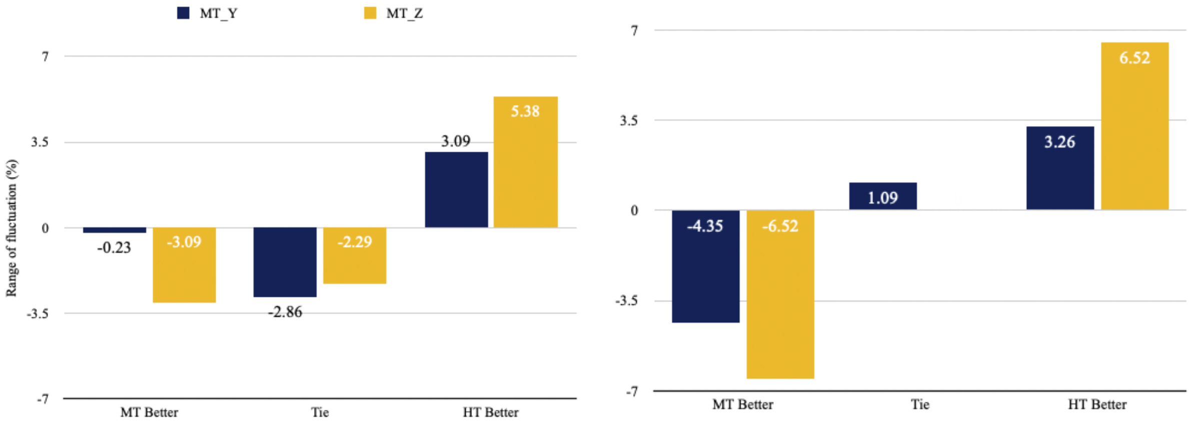
Human Parity According to Läubli et al. (2018), human parity is represented as a tie to HT, and super-

[2] https://www.kaggle.com/rmisra/news-category-dataset

[3] Note that the purpose of our evaluation does not involve clarification of what a better MT model is. In that context, we anonymize the models.

[4] www.dqf.taus.net

[5] http://www.taus.net



(a) $n = 874$

(b) $n = 184$

Figure 2: Score variation of Sign Test (BTS→ATS) for MT_Y and MT_Z.

human performance, to MT-better. We are concerned, however, that a tied rank can also be interpreted as a personal characteristic of indecisiveness. In that sense, we regard human parity as a sum of MT-better and tie ranks. It is to note that such a definition is only valid in the current study.

Annotator We collect judgments from 15 in-house researchers in the field of MT. They are native Koreans with good English proficiency. Some of them have previous experience in NMT evaluation. The participants are not previously informed of the involvement of HT in the task. Each is assigned to 1 - 2 HITs.

Inter-Annotator Agreement (IAA) To guarantee the reliability of the rankings, we calculate pairwise inter-annotator agreement with Cohen's kappa coefficient (K) (Cohen, 1960) following the precedence of WMT's ranking evaluation conducted until 2016 (Bojar et al., 2016).

5. Result

5.1. RR: Significance Test

Figure 2 displays a range of score variations from BTS to ATS of RR between MT_Y and MT_Z. The exact score is provided in Appendix A. A consistent tendency is that the proportion of HT-better is slightly increased in ATS ($y = 3.09pp$, $z = 5.38pp$) while that of MT-better and Tie decreases. When MT-better and Tie scores are combined, 3.09 and 5.38 percentage points of the judgments on human parity in MT_Y and MT_Z respectively turn out to be invalid. When narrowing down the scope to the contaminated sentences ($n = 184$), to our surprise, the tendency gets more intensified, as in Figure 2 (b). The loss of human parity claims reaches more than double in MT_Z ($-3.09\% \rightarrow -6.52\%$). We also notice that the percentage of Ties in both systems remains almost identical ($y = 1.09\%$, $x = 0\%$).

The fact that the score difference per category is more apparent in MT_Z than in MT_Y is also intriguing in that there is a definite possibility that a final ranking of the candidate systems can change. Such chance, however, is scarce in our case because more significant drop of MT-better or Tie occurs with MT_Z, which has seemingly lower performance than its counterpart.We confirm our hypothesis from the given result that the current test set of WMT 20 is more favorable to MT and leads to a false human parity achievement.

Meanwhile, the absolute RR scores of the three candidates are given in Table 7 and 8 in Appendix A. Figure 3 shows a score variation of the absolute ranking score of the three candidates when in BTS and ATS. Although the comparative rank is maintained identically in the order of [HT-MT_Y-MT_Z], the absolute score of HT increases up to 3.3% while that of MT descreases up to 4.67% (in MT_Z). Consequently, the gap between the two NMT models becomes more prominent. From such findings, we assume that systems behave differently with the contaminants of the test set. In our case, MT_Z is more vulnerable to them.

5.2. IAA

Figure 4 shows the IAA scores per group computed by Cohen's K. Each group is composed of two annotators who have been assigned to the same HIT. The average K score is 0.31.

6. Additional Work

6.1. Automatic Evaluation

While the influence of the contaminants on human parity is clarified, we get curious whether it is also observable in an automatic MT evaluation despite the small size of the data set ($n = 437$). BLEU (Papineni et al., 2002), TER, and chrF2 (Popović, 2015) are computed on BTS and ATS with Sacrebleu (Post, 2018). The sentences are tokenized with MeCab (Park and Cho, 2014) before the computation. We compare the result to Google Translate (GT) as a benchmark model

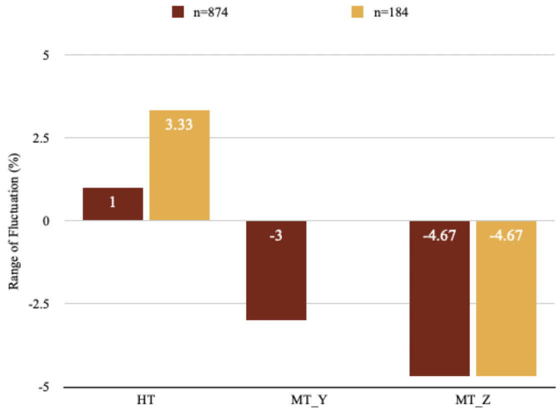


Figure 3: Absolute score variation of RR (BTS→ATS) in two volumes of the data set.

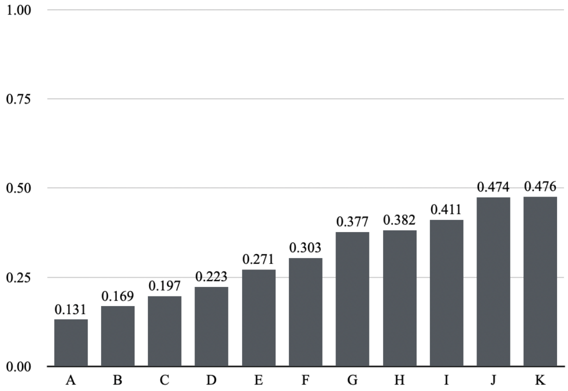


Figure 4: Inter-annotator agreement measured by 11 groups(A-K) by Cohen K. The scores are in ascending order.

to guarantee the compatible performance of the two anonymous models.[6] Table 4 shows that such changes barely affect the suggested metrics except for GT. The overall scores for GT have degraded in all three metrics.

All in all, more investigation is required in this regard based on the following observations: i) different from the outcome of RR (Section 5.1), MT$_Z$ obtains a better result; its correlation with human evaluation, thus, is uncertain, ii) the reliability of the automatic metrics in the Korean language is still dubious (Kim and Ventura, 2020), and iii) such marginal gap between BTS and ATS cannot be a shred of valid evidence.

	BTS			ATS		
	MT$_Y$	**MT$_Z$**	**GT**	**MT$_Y$**	**MT$_Z$**	**GT**
BLEU	19.78	20.19	13.12	19.78	20.24	9.49
TER	0.65	0.66	0.76	0.65	0.66	0.82
chrF2	0.27	0.27	0.19	0.27	0.27	0.16

Table 4: Result of automatic metrics for MT$_Y$, MT$_Z$, and Google Translate (GT) in BTS and ATS.

6.2. Qualitative Analysis

We suppose that sentence-level TER scores between BTS and ATS would hint at our interest in finding cases that have been critically influenced by the technical contaminants. Three sentences have scored above 0.8 in either MT$_Y$ or MT$_Z$, two of which have had contaminants in the source text. Taking a look at one of those examples ($MT_Z = 0.86$) in Figure 5 with back-translations of Google Translate (KO→EN), the contaminant in BTS-ST is an absence of an end quote. Just by adding it, we have obtained quite different system translations (in ATS-Y and ATS-Z). As expressed in the back-translations, ATS-Z's translation has lost a good deal of the source content, such as "*oh my*" and "*it's a lot of work*." A more interesting finding is their ranking scores. In BTS, the ranking of [*HT, MT$_Y$, MT$_Z$*] is either $[r = 1, y = 2, z = 1]$ or $[r = 1, y = 2, z = 2]$. In ATS, however, the score is converted into either $[r = 1, y = 1, z = 2]$ or $[r = 1, y = 2, z = 3]$. It is still unclear why such phenomenon happens, but we confirm that such minor changes in the source sentence can produce a very different translation, which affects to the human evaluation.

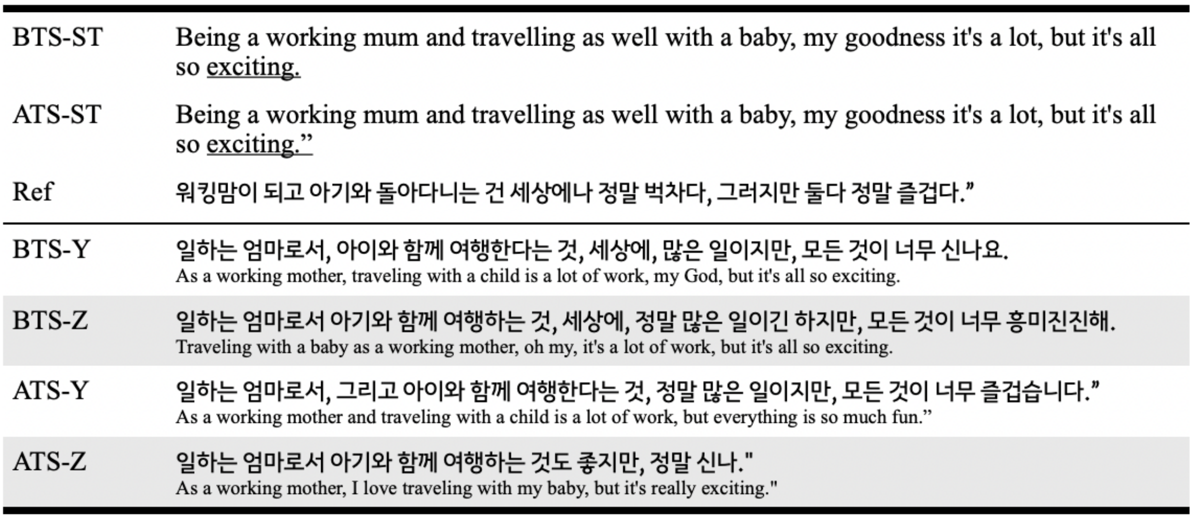
7. Conclusion

Many of us probably agree that a test set can have errors. Some would even say that it represents a real-world scenario and that it is acceptable. We, however, give proof that it is not satisfactory anymore if the human-machine parity of high-performing MT models is involved.

Technically, we identify nine types of contaminants and point out quotation marks as a primary culprit of the error. While doing so, we also confirm with the help of topic classification that the topic of the WMT 20 test set is heavily tilted toward world news and politics, while art is hardly visible. We show that such topical imbalance ignores the composition of its training set and disqualifies itself as a testbed. The in-depth study in this regard is left for a future work.

To verify the influence of technical contaminants on human-machine parity, we conduct an RR evaluation on two test sets (BTS and ATS) comparing two NMT systems and HT. We report that when Sign Test is concerned, the two MTs have lost about 5% of human parity claims in the clean test set (ATS) and that such tendency gets much more substantial when contaminated sentences are tested only. The sentence-wise TER scores show that system translations could

[6] Google Translate is known to be one of the most robust online translation models in many language pairs. Despite its fame, is was not employed in our experiment after having tested that it's false positive ratio of sentence-wise TER scores reached 88.69%, which meant that the contaminants were not a sole factor of the variation.



BTS-ST	Being a working mum and travelling as well with a baby, my goodness it's a lot, but it's all so exciting.
ATS-ST	Being a working mum and travelling as well with a baby, my goodness it's a lot, but it's all so exciting.”
Ref	워킹맘이 되고 아기와 돌아다니는 건 세상에나 정말 벅차다, 그러지만 둘다 정말 즐겁다.”
BTS-Y	일하는 엄마로서, 아이와 함께 여행한다는 것, 세상에, 많은 일이지만, 모든 것이 너무 신나요. As a working mother, traveling with a child is a lot of work, my God, but it's all so exciting.
BTS-Z	일하는 엄마로서 아기와 함께 여행하는 것, 세상에, 정말 많은 일이긴 하지만, 모든 것이 너무 흥미진진해. Traveling with a baby as a working mother, oh my, it's a lot of work, but it's all so exciting.
ATS-Y	일하는 엄마로서, 그리고 아이와 함께 여행한다는 것, 정말 많은 일이지만, 모든 것이 너무 즐겁습니다.” As a working mother and traveling with a child is a lot of work, but everything is so much fun.”
ATS-Z	일하는 엄마로서 아기와 함께 여행하는 것도 좋지만, 정말 신나." As a working mother, I love traveling with my baby, but it's really exciting."

Figure 5: MT$_Z$'s exemplary sentence of $TER = 0.86$. The contaminant is an absence of end quote (Category: Quotation marks). Back-translations are created from Google Translate. The source texts of BTS and ATS are brifed as BTS-ST and ATS-ST.

be edited up to 86% when the contaminants are revised. When qualitatively approached, we confirm that ranking judgments on that sentence become unfavorable. In the meantime, the side-effects of the contaminants on automatic evaluation metrics are questioned as an additional work, but it seems minor at this moment.

The current study is limited to a single language pair that has not been employed in WMT. We also acknowledge that such findings should be further examined on a larger scale with more resourceful language pairs and with publicly available NMT models. However, we believe that the WMT evaluation surroundings should be consistent at all times. With the results at hand, we cannot help but question the true goal of the news track at WMT. Is the real-world scenario what we want indeed? If we are to assess the maximum performance of NMT in comparison with human performance, we insist that the test set be technically impeccable and topically multifaceted to secure an unbiased evaluation experiment.

Acknowledgements

We would like to thank each and every one of our in-house researchers who put their time and efforts on our evaluation. We also appreciate Graham Neubig and Samuel Läubli for their valuable comments.

Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November. Association for Computational Linguistics.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. *CoRR*, abs/1608.04631.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August. Association for Computational Linguistics.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October. Association for Computational Linguistics.

doi:10.20944/preprints202110.0199.v1

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

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

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Edunov, S., Ott, M., Ranzato, M., and Auli, M. (2019). On the evaluation of machine translation systems trained with back-translation. *CoRR*, abs/1908.05204.

Graham, Y., Haddow, B., and Koehn, P. (2019). Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Graham, Y., Haddow, B., and Koehn, P. (2020). Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online, November. Association for Computational Linguistics.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Kim, A. and Ventura, C. C. (2020). Human evaluation of nmt & annual progress report: A case study on spanish to korean. *Revista Tradumàtica: tecnologies de la traducció*, 18.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In John A. Carroll, et al., editors, *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? A case for document-level evaluation. *CoRR*, abs/1808.07048.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., and Toral, A. (2020). A set of recommendations for assessing human-machine parity in language translation. *CoRR*, abs/2004.01694.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Park, E. L. and Cho, S. (2014). Konlpy: Korean natural language processing in python. In *Proceedings of the 26th Annual Conference on Human Cognitive Language Technology*, Chuncheon, Korea, October.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. *CoRR*, abs/1808.10432.

Toral, A. (2020). Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. *CoRR*, abs/2005.05738.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

A. Appendix

	MT$_Y$		**MT**$_Z$	
	BTS	ATS	BTS	ATS
MT better	117	115	116	89
Tie	246	221	180	160
HT better	511	538	578	625

Table 5: Sign test of MT$_Y$ and MT$_Z$ ($n = 874$, p-value $\leq$0.001).

	MT$_Y$		**MT**$_Z$	
	BTS	ATS	BTS	ATS
MT better	38	30	32	20
Tie	40	42	36	36
HT better	106	112	116	128

Table 6: Sign test of MT$_Y$ and MT$_Z$ ($n = 184$, p-value $\leq$0.001).

	HT	**MT**$_Y$	**MT**$_Z$
BTS	2.77	2.23	2.08
ATS	2.80	2.14	1.94

Table 7: Absolute score of RR. The total score is 3 (n=874).

	HT	**MT**$_Y$	**MT**$_Z$
BTS	2.68	2.20	2.14
ATS	2.78	2.20	2.00

Table 8: Absolute score of RR. The total score is 3 (n=184).