Assessing Human-Parity in Machine Translation on the Segment Level

Yvette Graham
ADAPT
Trinity College Dublin
ygraham@tcd.ie

Christian Federmann
Microsoft Research
chrife@microsoft.com

Maria Eskevich
CLARIN ERIC
Utrecht
maria@clarin.eu

Barry Haddow
School of Informatics
University of Edinburgh
bhaddow@inf.ed.ac.uk

Abstract

Recent machine translation shared tasks have shown top-performing systems to tie or in some cases even outperform human translation. Such conclusions about system and human performance are, however, based on estimates aggregated from scores collected over large test sets of translations and so leave some remaining questions unanswered. For instance, simply because a system significantly outperforms the human translator on average may not necessarily mean that it has done so for every translation in the test set. Furthermore, are there remaining source segments present in evaluation test sets that cause significant challenges for top-performing systems and can such challenging segments go unnoticed due to the opacity of current human evaluation procedures? To provide insight into these issues we carefully inspect the outputs of top-performing systems in the recent WMT19 news translation shared task for all language pairs in which a system either tied or outperformed human translation. Our analysis provides a new method of identifying the remaining segments for which either machine or human perform poorly. For example, in our close inspection of WMT19 English to German and German to English we discover the segments that disjointly proved a challenge for human and machine. For English to Russian, there were no segments included in our sample of translations that caused a significant challenge for the human translator, while we again identify the set of segments that caused issues for the top-performing system.

1 Introduction

Recent results of machine translation evaluation shared tasks indicate that state-of-the-art is now achieving and possibly even surpassing human performance, with the most recent annual Conference on Machine translation (WMT) news task providing extensive human evaluation of systems, concluding that several systems performed on average as well as human for English to Russian, English to German and German to English translation and a top system even surpassed human performance for the last two language pairs.

Since 2017 the official results of the WMT news tasks have been based on the human evaluation methodology known as Direct Assessment (DA) (Graham et al., 2016), due to its many advantages over older technologies. DA, for example, includes quality control mechanisms that allow data collected anonymously from crowd-sourced workers to be filtered according to reliability.[1] Although WMT news task results are admittedly based on substantially more valid methodology than those usually found in general in system comparisons using automatic metrics such as BLEU, results in WMT human evaluations still leave some questions unanswered. For example, DA scores are based on average ratings attributed to translations sampled from large test sets, and although such methodology does allow application of statistical significance testing to identify potentially meaningful differences in system performance, they do not provide any insight into the reasons behind a significantly higher score or the degree to which systems perform better when translating individual segments. Furthermore, DA score distributions produced in the human evaluation of the news task are based on individual DA scores that alone cannot be relied upon to reflect the quality of individual segments (Graham et al., 2015).

Past work, has however provided a means of running a DA human evaluation in such a way that DA scores accurately reflect the performance of a system on a given individual segment (Graham

[1] DA is also used in other task evaluations such as Video Captioning and Multilingual Surface Realisation (Awad et al., 2019; Graham et al., 2018; Mille et al., 2019).

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4199–4207
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

et al., 2015). This method comes with the trade-off of requiring substantially more repeat assessments per segment than the test set level evaluation generally run, for example, to evaluate all primary submissions in the WMT news task. In this work we demonstrate how this method has the potential to be employed as a secondary method of evaluation in WMT tasks for a smaller subset of systems to provide segment-level insight into why the top-performing systems outperform one another or indeed to investigate the degree to which human and machine performance differs for individual segments.

2 Related Work

Over the past number of years, machine translation has been biting at the heels of human translation for a small number of language pairs. Beginning with the first claims that machines have surpassed human quality of translation for Chinese to English news text, conclusions received with some skepticism and even controversy (Hassan et al., 2018), as claims of human performance resulted in re-evaluations that scrutinized the methodology applied, highlighting the influence of reverse-created test data and lack of wider document context in evaluations (Läubli et al., 2018; Toral et al., 2018). Despite re-evaluations taking somewhat more care to eliminate such sources of inaccuracies, they additionally included some potential issues of their own, such as employing somewhat outdated human evaluation methodologies, non-standard methods of statistical significance testing and lack of planning evaluations in terms of statistical power. Graham et al. (2019, 2020), on the other hand re-run the evaluation, identify and fix remaining causes of error, and subsequently confirm that, on the overall level of the test set, with increased scrutiny on evaluation procedures, conclusions of human parity were still overly ambitious at that time.

It was not long before results were shown to have reached human performance however, according to more scrutinous human evaluation procedures, as one year later at WMT 2019, MT system performance for some language pairs reached human performance and even surpassed it for two language pairs (Barrault et al., 2019).

Although the admittedly rigorous human evaluation employed in WMT evaluations provides valid conclusions about systems significantly outperforming human translation, it nonetheless employs the somewhat opaque average Direct Assessment scores computed over large test sets of segments that subsequently leave some important questions unanswered in terms of human parity. For example, even if a system performs better on average than a given human translator, this does not necessarily mean that the system translates every sentence better than the human translator. When a tie occurs between human and machine translation, it would be useful to know how performance compares between the two on individual segments. The current WMT human evaluation methodology does not allow for this, however.

In this paper, we carry out fine-grained segment-level comparison of system and human translations using human evaluation and provide a comparison on the segment-level of the top-performing MT systems from WMT-19 news task and the human translator for all language pairs in which a system was shown to either tie (English to Russian) or surpass human performance (English to German; German to English). Human evaluation is required, as opposed to segment-level BLEU, for example, because metrics such as BLEU are not sufficiently accurate to identify fine-grained segment-level differences in quality, as can be seen from low correlations with human assessment (Ma et al., 2019). We make all code and data collected in this work publicly available to aid future research.[2]

3 Segment-level Direct Assessment

Segment-level Direct Assessment requires running human evaluation with sampling of translations carefully structured to ensure that repeat assessment of the same set of translations occurs a minimum of 15 times for both the translations produced by the systems of interest (Graham et al., 2015). For example, this can be carried out for a reduced number of translations and for a reduced number of systems than the entire test set, since collecting 15 repeat assessments makes exhaustive segment-level evaluation for every participating system likely to be overly costly. It is reasonable to focus the segment-level evaluation on a sample of approximately 500 translations selected at random for the two top-performing systems or indeed, as we do now, the top-performing system and the human translator. An important consideration however, is that regardless of which systems may

[2] http://www.scss.tcd.ie/~ygraham/emnlpfindings20datacode

be selected for fine-grained segment-level analysis, segment-level evaluation should be run for precisely *the same set of segments for all systems of interest* so that a comparison of the performance of systems on the same segments will ultimately be possible.

The desired number of source language segments should therefore be sampled at random from the test set before pooling target side translations for the systems of interest, shuffling and arranging them within human intelligence tasks (HITs). We construct HITs of 100 translated segments, subsequently evaluated by humans blind to which system has produced each translation, or as in our case, blind to whether a human or machine produced the translated segment. The configuration of DA we employ is a source-based segment-level evaluation in which human assessors are (i) shown the source language input segment; (ii) the translated text (either human or machine-produced); and (iii) asked to rate the adequacy of the translation on a 0–100 rating scale. Source-based DA has the advantage of freeing up reference translations so that they can be included in the evaluation as if they had been produced by a system. Source-based DA comes with the trade-off, however of requiring bilingual human assessors.

4 Experiments

In order to investigate the degree to which human and machine perform differently on individual test set segments, we run segment-level DA on translations of the same random sample of 540 segments by the top-performing system and the human translator. We do this for each language pair in which there was a tie with human performance WMT-19 (English to Russian) or where machine translation performance had surpassed human translation quality (German to English; English to German).[3]

In order to access the bilingual speakers required for the source-based DA configuration we run all source-based DA HITs on an in-house crowdsourcing platform. In total 108,829 assessments were collected via the in-house platform. After removing quality controls, we ended up with 87,211 assessments for which we are confident of worker reliability, and employ all those assessments in our final analysis.

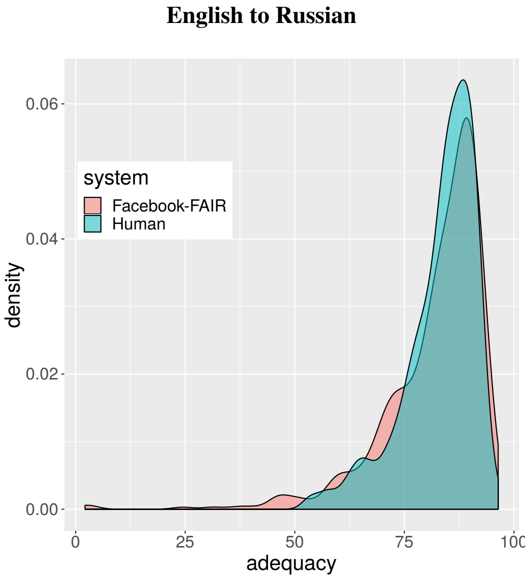

Figure 1: Density plot of sample of 540 accurate segment-level DA scores for English to Russian news translation for top-performing system, FACEBOOK-FAIR, in WMT-19 versus the human translator where in the official results the system tied with human performance; Human denotes evaluation of segments translated by the creator of the standard WMT reference translations

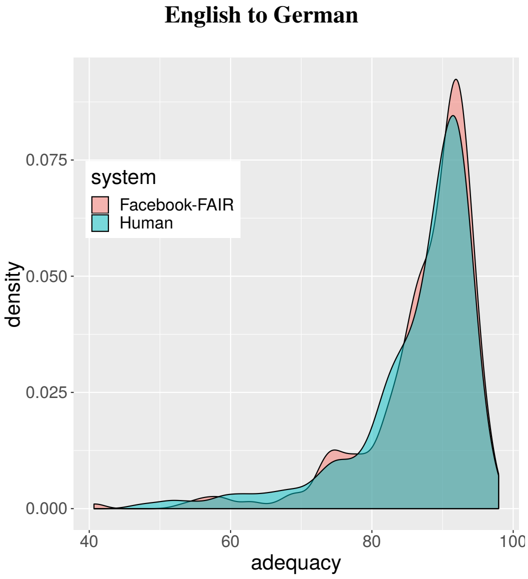

Figure 2: Density plot of sample of 540 accurate segment-level DA scores for English to German news translation for the top-performing system, FACEBOOK-FAIR, in WMT-19 versus the human translator where in the official results the system beat human performance; Human denotes evaluation of segments translated by the creator of the standard WMT reference translations

[3] For German to English translation, although HUMAN and the top-performing system, FACEBOOK-FAIR, are ranked in the same cluster, the system significantly outperforms human translation in head-to-head significance test results.

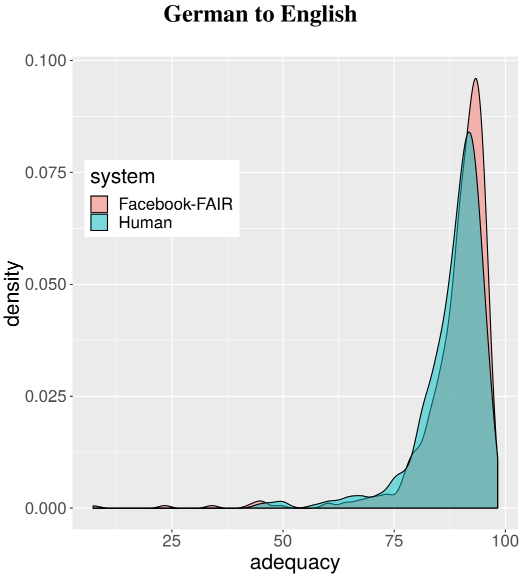

Figure 3: Density plot of sample of 540 accurate segment-level DA scores for German to English translation new translation for the top-performing system, FACEBOOK-FAIR, in WMT-19 versus the human translator where in the official results the system beat human performance; Human denotes evaluation of segments translated by the creator of the standard WMT reference translations

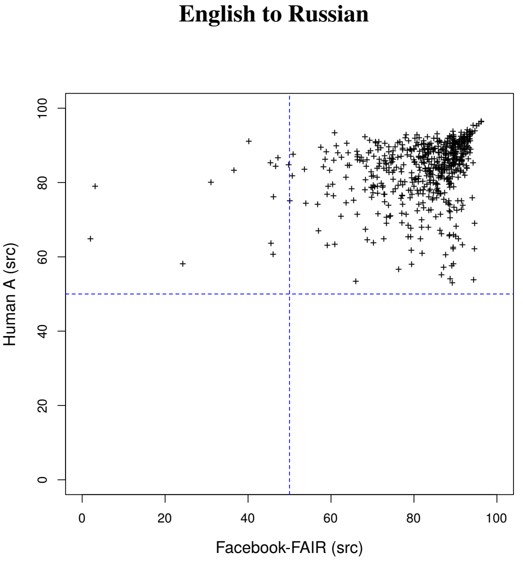

Figure 4: Scatter plot of accurate segment-level DA scores for top-performing system, FACEBOOK-FAIR in WMT-19 versus the human translator where in the official results the system tied with human performance; Human A denotes evaluation of segments translated by the creator of the standard WMT reference translations; src denotes a source-based configuration of Direct Assessment was employed to collect scores; segment-level scores for human and machine are the average of a minimum of 15 human assessment scores

Figures 1, 2 and 3 include density plots for human translation and the top-performing FACEBOOK-FAIR system (Ng et al., 2019) for the same 540 translated segments from WMT-19 for the three language pairs we investigate.

For German to English and English to German translation in Figures 2 and 3 a similar pattern emerges in terms of comparison of human and machine-translated segments, as for both a slightly larger proportion of FACEBOOK-FAIR translations are scored high compared to the human translator – as can be seen from the higher red peak close to the extreme right of both plots indicating that the machine produces a marginally higher number of translations with higher levels of adequacy. For English to Russian translation, however, a different pattern occurs, as shown in Figure 1, as it appears that there are locations lower down on the adequacy scale in which the FACEBOOK-FAIR system performs worse than the human translator in three noticeable locations within its score distribution. However, these differences between language pairs are somewhat unsurprising considering that human and system were tied for English to Russian but system beat human in terms of statistical significance for both English to German and German to English.

4.1 Human V FACEBOOK-FAIR: English to Russian

As revealed in the WMT-19 human evaluation results, a single system achieved a statistical tie with human assessment for English to Russian news translation. Differences in average overall scores computed on large test sets still leave some questions unanswered however, particularly in terms of which specific source inputs the machine or even human translator might still find challenging. Furthermore it does not provide any insight into differences in performance for specific source language input segments.

Since we desire the ability to examine differences in translations of individual source segments for machine and human we examine scatter plots of accurate segment scores for translations of the same input source segment by the human translator and the top-performing machine, shown in Figure 4 for English to Russian for WMT-19 data.

For English to Russian translation, the scatter plot of adequacy scores for human translator versus machine shown in Figure 4, in which each "+" signifies the translation of the same source language input test segment, reveals distinct levels of performance for human versus machine for individual segments. Figure 4 reveals that as expected the vast majority of translations score high for both human and machine translations, depicted by the location of the main bulk of translations within the upper right quadrant, as both human and machine translations in this quadrant received an average score above 50%. A perhaps more interesting insight revealed by Figure 4 is the lack of translations appearing in the bottom right quadrant and this indicates that when the system does well on an input source segment so does the human. The reverse cannot be said however of the system, as the upper left quadrant in Figure 4 for English to Russian contains albeit a relatively small number of segments (12 or 2.4%) for which Facebook-FAIR translates poorly while corresponding adequacy scores for the human translator remain above 50%.

To gain more insight into what might take place in the case that either the machine or human performs poorly for the input segments scored below the 50% threshold for English to Russian translation see Table 1, where we include the full translation examples for the two lowest scored FACEBOOK-FAIR translations. In example (a) in Table 1 the system is scored lower because it translates *an unknown person on Capitol Hill* incorrectly. While the human translator correctly expresses the fact that the person is from *Capitol Hill*, the system instead implies that the unknown person is *on Capitol hill*, i.e. as if that person were physically standing on a hill. All the other differences between the human and machine in terms of selection of words in the Russian translation are not critical and read well in terms of the fluent Russian.

In example (b) in Figure 1 there is firstly a mistake in the system translation as it translates *detained* into Russian as *delayed* instead of the correct translation that is produced by the human translator. Secondly, in this same example, the system translates *migrant children* using a Russian term that only refers to children who are migrants themselves, while the human translator uses an arguably better term that includes both children who are migrants and the children of migrants. Finally, in example (b) the system translation appears to lose the intensity of the causality implication that the sentence originally has in English, while the human translation keeps this using the active form of the verb. Remaining English to Russian translations for which the system score falls below 50% are included in Appendix A.

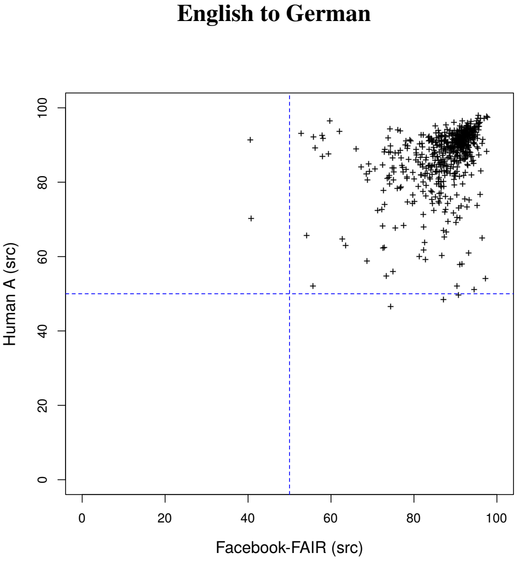

Figure 5: Scatter-plot of segment-level DA scores for top-performing system, FACEBOOK-FAIR in WMT-19 versus human translator; Human A (src) denotes evaluation of segments translated by the creator of the standard WMT reference translations in a source-based configuration of DA; segment-level scores for human and machine are the average of a minimum of 15 human assessment scores

As mentioned previously, for this English to Russian our analysis found no translations for which the human translator performed very poorly while the system succeeded.

4.2 Human V Super-human FACEBOOK-FAIR: English to German

In the official WMT-19 human evaluation results of the English to German news task, again the same single system, FACEBOOK-FAIR, stood out as quite remarkably outperforming the human translator according to human assessment scores computed over the entire test set (Barrault et al., 2019). In order to further investigate this super-human performance, after collecting accurate segment-level scores for translations of the same 540 source language input segments for both FACEBOOK-FAIR and the human translator, we plot corresponding adequacy scores in Figure 5.

In contrast to English to Russian (Figure 4), and perhaps not surprisingly since the system significantly outperforms the human translator as opposed

			DA (%)
(a)	Source :	*The information appeared online Thursday, posted by an unknown person on Capitol Hill during a Senate panel's hearing on the sexual misconduct allegations against Supreme Court nominee Brett Kavanaugh.*	
	Facebook-FAIR:	Информация появилась в сети в четверг, размещенная неизвестным лицом на Капитолийском холме во время слушаний в сенатской комиссии по обвинениям в сексуальных домогательствах в адрес кандидата в Верховный суд Бретта Кавано.	3.3
	Human:	Информация появилась онлайн в четверг, опубликованная неизвестным с Капитолийского холма во время слушаний коллегии Сената по поводу обвинений в неподобающем поведении против кандидата в Верховный суд Бретта Кавано.	78.9
(b)	Source :	*The number of detained migrant children has spiked even though monthly border crossings have remained relatively unchanged, in part because harsh rhetoric and policies introduced by the Trump administration have made it harder to place children with sponsors.*	
	Facebook-FAIR:	Число задержанных детей-мигрантов резко возросло, несмотря на то, что ежемесячные пересечения границы остались относительно неизменными, отчасти из-за жесткой риторики и политики, введенной администрацией Трампа, стало труднее помещать детей к спонсорам.	2.1
	Human:	Число задержанный детей мигрантов резко подскочило, хотя среднемесячное количество переходов границы осталось относительно неизменным, частично потому, что жёсткая риторика и политика, принятые администрацией Трампа, усложнили задачу найти детям спонсоров.	64.8

Table 1: English to Russian example translations from WMT-19 news task for which the top-performing system performed poorly; DA denotes average direct assessment scores for translations computed on a minimum of 15 human assessments; DA scores below the 50% threshold highlighted in orange; DA scores above the 50% threshold highlighted in blue

to merely tying with it, the English to German system shows fewer machine translations receiving a low adequacy score combined with a high human score, as only two translations appear in the top-left quadrant of Figure 5. This highlights the fact that even though on average the system performs incredibly well, by on average outperforming human translation, there remains the possibility that this can take place in combination with a albeit small number of poor translations.

To gain more insight into what might take place in the case that either the machine or human performs poorly for the input segments scored below the 50% threshold see Table 2. Two out of the five translations that scored below 50% by either human or machine were translated worst by machine as opposed to the human translator as can be seen by the lower DA scores (a) and (b) in Table 2. Firstly, in example (a) in Table 2 the system translation deviates from the syntactic structure of the source input sentence. It additionally ignores *and* in addition to translating *scene* as *Unfallort* (lit: *location of the accident*). In contrast, the human translator instead produces *Ort des Geschehens* which is arguably a better way to express *scene*.

In example (b) in Table 2, the source word *trough* is mistranslated as *Trog* by the system, which is a more common translation of the word *trough* but is in this context an incorrect lexical choice given that the source input sentence originates in the weather report domain, for which *Tief* is the appropriate translation, which the human translator correctly translates.

Despite the system performing poorly on two segments for which the human translates correctly, perhaps more surprising is that there are three source input segments for which the machine translates well but the human translator does not. In example (c) in Table 2, the human translates *broadcast networks* somewhat too literally as *Rundfunknetze* instead of *Rundfunksender*. In addition, the human translator incorrectly changes the tense. Finally in example (c) in Table 2 *full Senate* is again translated too literally into *vollem Senat*.

In example (d) in Table 2, the human translator chooses the incorrect present tense for the main verb, *kündigt ... an* as opposed to the future tense. Lastly, in example (e) in Table 2, the human translator converts *two-foot* into *60 cm* which is only approximately correct, the source word *brim* is translated into *Rand* which is arguably correct but is nonetheless an unusual lexical choice compared to the system translation, *Krempe*. Again, tense in the latter part of the source input sentence is not preserved well in the human translation.

			DA (%)
(a)	Source :	*The driver of the car stopped and paramedics attended, but the man died at the scene.*	
	Facebook-FAIR:	Der Fahrer des Autos hielt an, Sanitäter kümmerten sich um ihn, doch der Mann starb noch am Unfallort.	40.7
	Human :	Der Fahrer des Autos hielt an und Sanitäter kamen, aber der Mann starb am Ort des Geschehens.	91.3
(b)	Source :	*The approaching trough will bring some locally heavy rain to parts of the Southern California coastline.*	
	Facebook-FAIR:	Der herannahende Trog wird Teilen der südkalifornischen Küste lokal heftigen Regen bringen.	40.9
	Human :	Das sich nähernde Tief wird einige örtlich starke Regenfälle für Teile der südkalifornischen Küste mit sich bringen.	70.1
(c)	Source :	*The cable and broadcast networks were all covering live hours later, when the Judiciary Committee was to vote to advance Kavanaugh's nomination to the full Senate for a vote.*	
	Facebook-FAIR:	Die Kabel- und Rundfunksender berichteten alle live Stunden später, als der Justizausschuss abstimmen sollte, um Kavanaughs Nominierung dem gesamten Senat zur Abstimmung vorzulegen.	74.5
	Human :	Die Kabel- und Rundfunknetze haben später live übertragen, als der Justizausschuss abstimmen sollte, um die Ernennung von Kavanaugh zum vollen Senat zur Abstimmung voranzutreiben.	46.5
(d)	Source :	*Foreign buyers are set to be charged a higher stamp duty rate when they buy property in the UK - with the extra cash used to help the homeless, Theresa May will announce today.*	
	Facebook-FAIR:	Ausländischen Käufern soll beim Kauf von Immobilien in Großbritannien eine höhere Stempelsteuer in Rechnung gestellt werden – mit dem zusätzlichen Geld, das für Obdachlose verwendet wird, wird Theresa May heute bekannt geben.	90.9
	Human :	Ausländischen Käufern wird beim Kauf von Immobilien in Großbritannien ein höherer Stempelsteuersatz in Rechnung gestellt - das zusätzliche Geld wird für Obdachlose verwendet werden, kündigt Theresa May heute an.	49.5
(e)	Source :	*The out-sized hats come hot on the heels of 'La Bomba', the straw hat with a two-foot wide brim that's been seen on everyone from Rihanna to Emily Ratajkowski.*	
	Facebook-FAIR:	Die überdimensionalen Hüte sind auf den Fersen von "La Bomba", dem Strohhut mit zwei Fuß breiter Krempe, den man von Rihanna bis Emily Ratajkowski gesehen hat.	87.3
	Human :	Die überdimensionalen Hüte haben sich an die Fersen von "La Bomba" geklebt, dem Strohhut mit einem 60 cm breiten Rand, der bei jedem von Rihanna bis Emily Ratajkowski zu sehen ist.	48.3

Table 2: English to German translations from WMT-19 news task for which either the top-performing system or human translator perform poorly; DA denotes average direct assessment scores for translations computed on a minimum of 15 human assessments; DA scores below the 50% threshold highlighted in orange; DA scores above the 50% threshold highlighted in blue

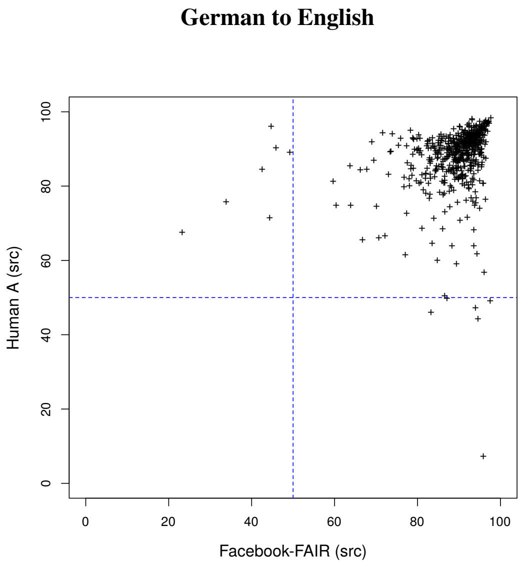

Figure 6: Scatter plot of adequacy scores of translations of the same source language input segment produced by (i) human and (ii) top-performing machine translation system from WMT-19, FACEBOOK-FAIR for German to English where machine significantly outperformed human translation

4.3 Human V Super-human FACEBOOK-FAIR: German to English

For German to English translation, the scatter-plot of translation scores for our 540 source segment sample shown in Figure 6 reveals the bulk of translations located to a more extreme degree in the upper right corner of the plot compared to the other two language pairs. Like both English to Russian and English to German, there are segments for t his language pair for which the top-performing system, FACEBOOK-FAIR performs poorly on compared to the human translator, as seven source segments (1.4%) appear in the upper-left quadrant, where the system received an adequacy score lower than 50% while the human translation received a score higher than 50%. Like English to German, however, for German to English translation, the reverse is also true, there are translations that catch out the human translator, for which he/she received a low score, while for the same source input, the machine receives a high score. Such translations, there are six in total (1.2%), are located in the bottom-right quadrant of Figure 6.

Table 3 shows the most extreme examples in

			DA (%)
(a)	Source :	*Im Ziel warf er sein Paddel vor Freude weg und reckte beide Arme siegessicher in die Höhe - wohlwissend, dass es mindestens für eine Medaille reichen würde.*	
	Facebook-FAIR:	At the finish, he threw away his paddle for joy and raised both arms in victory - knowing that it would be enough for at least one medal.	23.4
	Human :	He threw his paddle with joy at the finishing line and, confident of victory, threw both arms in the air - safe in the knowledge that his efforts would secure him a medal.	67.5
(b)	Source :	*Zur Vorsicht wurde auch noch der ÖAMTC-Notarzthubschrauber gerufen.*	
	Facebook-FAIR:	The ÖAMTC emergency medical helicopter was also called out as a precaution.	42.7
	Human :	As a precautionary measure, an emergency air ambulance helicopter was also called into action.	84.5
(c)	Source :	*Hintergrund ist Musks überraschende Ankündigung vom August, Tesla von der Börse nehmen zu wollen.*	
	Facebook-FAIR:	The background is Musk's surprise announcement in August that he would take Tesla off the stock market.	96.0
	Human :	The background is Musk's surprise announcement in August to take Tesla off the stock exchange.	7.3
(d)	Source :	*Zum 100-Jahr-Jubiläum der Republik, das in diesem Gedenkjahr seit mittlerweile fast zehn Monaten gefeiert wird, sind zahlreiche neue Bücher erschienen, die diese Frage meist im Rückblick auf die vergangenen hundert Jahre beantworten.*	
	Facebook-FAIR:	On the occasion of the 100th anniversary of the Republic, which has been celebrated in this commemorative year for almost ten months now, numerous new books have been published, most of which answer this question in retrospect of the past hundred years.	97.7
	Human :	At the 100-year anniversary of the republic that has been celebrated in this commemorative year for almost ten months, many new books appeared that answer this question mainly looking back over the past hundred years.	49.1

Table 3: German to English translations from WMT-19 news task for which either the top-performing system or human translator perform poorly; DA denotes average direct assessment scores for translations computed on a minimum of 15 human assessments; DA scores below the 50% threshold highlighted in orange; DA scores above the 50% threshold highlighted in blue

terms of contrast in adequacy scores for human versus machine translation for German to English for the top-performing system FACEBOOK-FAIR. Two of the examples (a) and (b) show segments for which the system performs worse than the human translator and on close inspection we can see why this could be. For example, the machine translates the source segment in example (a) in Table 3 too literally and omits the phrase "in the air". Although the human translator scores higher at 67.5% they are still docked some marks probably because the human translator has also slightly mistranslated the German verb *wegwerfen – to throw away*, omitting *away* from his/her translation. Example (b) in Table 3 the machine translation system is hindered by the presence of an unknown acronym containing the German umlaut that remains as such incorrectly present in the English translation, receiving a score of 42.7%. The human translator, achieving a score of 84.5%, handles this better by omitting the acronym from the translation, but still there is possibly some meaning missing from its translation.

Table 3 additionally includes some examples, (c) and (d), in which it was the human translator who was caught off guard by a particular source segment and substantially scored lower than the machine for its translation. For instance, example (c) in Table 3 the system correctly translates the German term *Börse* as *stock market* while the human translator chooses *stock exchange* which has likely caused a low human assessment score, as in general companies are added and removed from stock markets as opposed to stock exchanges. In example (d) in Table 3 it is likewise the human translator who translates the German term *erschienen* as appeared instead of the more appropriate *published* produced by FACEBOOK-FAIR. In addition the human misplaces the translation of *meist – most* – which refers back to the books in the preceding phrase and attaches it to the translation of *Rückblick – looking back* or *retrospect* – while the machine correctly translates *meist*. Remaining German to English translations for which either the system or human score falls below 50% are included in Appendices B and C.

5 Conclusions

The question we ask in this work is highly relevant – what are the differences between human translations and the top MT translations on the segment level when "human parity" is reached. For the English-Russian system our analysis makes it clear that there are a number of segments where the human did better than the machine, but on close inspection of these sentences, there appears to be no generalizable difference that clearly characterizes these kinds of sentences.

For English to/from German, the situation between human and machine is more finely balanced, and segment-level analysis has shown only a small number of random errors on each side, revealing only minor differences are present even on the segment level when we compare human and machine translations.

Acknowledgments

This study was supported by the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Trinity College Dublin funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund, and has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825299 (Gourmet). We would also like to thank the anonymous reviewers for their feedback.

References

George Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, A. F. Smeaton, Y. Graham, and W. Kraaij. 2019. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. In *Proceedings of TRECVID*, volume 2019.

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

Yvette Graham, George Awad, and Alan Smeaton. 2018. Evaluation of automatic video captioning using direct assessment. *PLOS ONE*, 13(9):1–20.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Virtual. Association for Computational Linguistics.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 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*, Denver, Colorado.

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. *CoRR*, abs/1803.05567.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Neural Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *EMNLP 2018*, Brussels, Belgium. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, OndÅ™ej Bojar, and Yvette Graham. 2019. Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, and Leo Wanner. 2019. The second multilingual surface realisation shared task (SR'19): Overview and evaluation results. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 1–17, Hong Kong, China. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 News Translation Task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *CoRR*, abs/1808.10432.