On the Stability of System Rankings at WMT

Rebecca Knowles
National Research Council Canada
Rebecca.Knowles@nrc-cnrc.gc.ca

Abstract

The current approach to collecting human judgments of machine translation quality for the news translation task at WMT – segment rating with document context – is the most recent in a sequence of changes to WMT human annotation protocol. As these annotation protocols have changed over time, they have drifted away from some of the initial statistical assumptions underpinning them, with consequences that call the validity of WMT news task system rankings into question. In simulations based on real data, we show that the rankings can be influenced by the presence of outliers (high- or low-quality systems), resulting in different system rankings and clusterings. We also examine questions of annotation task composition and how ease or difficulty of translating different documents may influence system rankings. We provide discussion of ways to analyze these issues when considering future changes to annotation protocols.

1 Introduction

At the WMT (now Conference on Machine Translation) shared task on news translation, research groups build machine translation systems to accurately translate news data, as tested on test sets of recent news documents. The systems are clustered and ranked on their performance as judged by human annotators. The way that human judgments of translation quality have been collected has varied over the course of WMT's history.

In this work, we examine how changes in the collection of human judgments over the last three years have resulted in rankings that are now less robust to the effects of outliers (high- or low-performing systems) and overall annotation task composition. We replicate the human judgment rankings from 2018-2020, perform simulations for reranking, and examine issues of annotation task composition and translation difficulty. We find that sampling sentences for annotators to annotate by document – intended as a step towards evaluating sentences in context – reintroduces a known problem from the earlier era of relative rankings, namely that systems suffer or benefit in their rankings based on the quality of the other data being rated alongside them in the same annotation tasks.

We begin with a discussion of the progression of direct assessment (DA) styles employed in WMT evaluations (§2) and how scoring is performed (§3), before delving into theoretical and practical understandings of the z-scores used to rank systems (§4 and §5), including simulations and analysis of specific examples. We also discuss issues around document distribution and translation difficulty (§6), and close with considerations for downstream impacts (§7) and future study (§8).

2 Historical Context

In 2016, WMT added direct assessment (DA) scoring of system outputs as an investigatory ranking, with relative ranking (RR) remaining the official scoring mechanism (Bojar et al., 2016). In relative ranking, five system outputs for a given segment were ranked in comparison to one another, from which pairwise translation comparisons were generated; these were then used to produce overall system rankings by means of the TrueSkill algorithm (Herbrich et al., 2007; Sakaguchi et al., 2014). Relative ranking can be used to compare systems, but does not provide an absolute score, thus obscuring how close a good system is to a "perfect" translation or, at the other extreme, how poor a system is as compared to others.

The following year, 2017, WMT adopted DA as its main assessment format on the basis of high Pearson correlations between RR and DA in the previous year's investigations (Bojar et al., 2017). In DA (Graham et al., 2013, 2014, 2016), annotators provide an absolute numerical score (0-100) for MT output adequacy (at the sentence level or at the document level) using a sliding scale.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 464–477
November 10–11, 2021.

The use of DA has changed since it was first introduced to WMT. In 2016, it was trialed for monolingual evaluations of translation fluency and monolingual evaluations of adequacy. Here we provide an overview of changes from the 2016 task to the present, based on Findings papers descriptions.

Bojar et al. (2016) noted that the 2016 version of DA assessments has the potential to avoid a known bias of the RR setup. In RR, each rating task consisted of ranking the outputs of five systems on the same input segment, and "a system may suffer more losses if often compared to the reference, and similarly it may benefit from being compared to a poor competitor" (Bojar et al., 2011). In the 2016 DA setup, translations were annotated in sets of 100, including quality assurance tasks, but each segment was annotated individually, rather than in direct comparison to other system output for the same segment.[1] Quality assurance tasks can include *references* (which should score highly), *"bad" references* (which should score poorly; these are produced by randomly replacing substrings in references to degrade quality), and *repeat assessments* of a segment (which should be scored consistently).

In 2017, DA was adopted as the main annotation style, with exact duplicate segment translations being able to be scored just once (rather than once per system that produced them) and with human assessment scores "standardized according to each individual human assessor's overall mean and standard deviation score" (Bojar et al., 2017).

Bojar et al. (2018) describes two setups for the 2018 DA tasks, a standard structure (with repeat pairs, "bad" references, and references, as quality assurance) and an alternate setup where an additional constraint was imposed, such that within each 100-translation task, for each input the task would include the corresponding output of *all* MT systems. This makes a tradeoff between the aim of DA (to make absolute score judgments rather than relative ones) and getting a single annotator to provide scores for all systems' output of the same source input (which risks reintroducing some form of relative judgement to the task). This is also the first year that the findings paper explicitly spells out the goal of the way tasks (referred to here using the Amazon Mechanical Turk nomenclature "Human Intelligence Task" or HIT) are built in the standard HIT structure:

> [...] within each 100-translation HIT, the same proportion of translations are included from each participating system for that language pair. This ensures the final dataset for a given language pair contains roughly equivalent numbers of assessments for each participating system. This serves three purposes for making the evaluation fair. Firstly, for the point estimates used to rank systems to be reliable, a sufficient sample size is needed and the most efficient way to reach a sufficient sample size for all systems is to keep total numbers of judgments roughly equal as more and more judgments are collected. Secondly, it helps to make the evaluation fair because each system will suffer or benefit equally from an overly lenient/harsh human judge. Thirdly, despite DA judgments being absolute, it is known that judges "calibrate" the way they use the scale depending on the general observed translation quality. With each HIT including all participating systems, this effect is averaged out.[2]

The 2018 shared task also introduced source-based DA, trialling a bilingual version of the task. Rather than scoring MT output against a reference, this version scores it against the source segment, which allows human references to be scored as a "human system" rather than solely as a QA task. They raise a number of potential cautions against drawing strong conclusions, namely that bilingual DA is not yet validated, the alternate task structure may introduce biases, the year's sample size for source-based DA was smaller than 1,500 judgments per system, and that there may be quality issues with some reference segments.

In 2019, WMT introduced additional versions of DA (Barrault et al., 2019). They used monolingual (reference-based) assessment for translation into English and for language pairs that did not include English at all. For translation out of English, they performed bilingual (source-based) DA. The style of DA used in previous years is renamed to SR–DC (Segment Rating without Document Context), as a new style, SR+DC (Segment Rating *with* Document Context) is introduced. In the new SR+DC style, the full translation of a single docu-

[1] It is still possible that there may be biases based on the segments observed in any given set of 100.

[2] Here we reproduce this quote from Barrault et al. (2019), though it appears consistent 2018-2020.

ment by a single MT system is shown to the annotator in order (but still scored segment-by-segment);[3] a task consists of multiple such documents. The generation of annotation tasks is described as follows: all documents translated by all systems are pooled, then sampled (without replacement) until up to 70 segments are selected, at which point quality control documents are added, and finally the order of documents in the task is shuffled. Barrault et al. (2020) uses both SR–DC and SR+DC styles.

3 Scoring

In order to experiment with questions surrounding human evaluation, it is necessary to understand and be able to replicate the official scores produced by WMT. For the human annotations of interest (segment-level evaluation, with or without document context), there are two main types of scores: raw scores and z-scores, with the latter used as the official ranking. These are presented in a table, ordered by z-score, and clusters of systems deemed statistically significantly different (according to a Wilcoxon rank-sum test $p < 0.05$) are separated by horizontal lines.[4]

Following the approach used at WMT, after removing any HITs deemed unacceptable due to quality issues, we calculate raw and z-scores for systems as follows. First, any worker ID whose scores have a standard deviation of 0 is removed. Given a raw score x generated by the worker with worker ID W, its corresponding z-score z is computed as

$$z = \frac{x - \text{mean}(y \in W)}{\text{std}(y \in W)} \quad (1)$$

where $\text{mean}(y \in W)$ is the mean of *all* raw scores generated by worker W, and $\text{std}(y \in W)$ is the standard deviation of *all* raw scores generated by that worker. When we say that the mean and standard deviation are computed from *all* raw scores from a given worker ID, this includes references (which are treated as systems in SR+DC but are treated as quality assurance in SR–DC), "bad references" (which are only ever used for quality assurance), and repeats.[5] However, after computing the mean and standard deviation, only a subset of scores are used to actually compute system averages: those with type "SYSTEM" or "REPEAT" (discarding "BAD_REF" and "REF" types).[6] To compute averages (raw or z-score), first an average is computed for any "SYSTEM" or "REPEAT" scores that share the same system ID, the same document ID, *and* the same sentence ID; that is, if a given sentence of a given document was annotated multiple times for a particular system, we first average those scores (so that more frequently annotated sentences do not receive more weight). Then, for each system, all of its "SYSTEM" or "REPEAT" type scores are averaged, resulting in a system-level score.

We note that the 2019 and 2020 document context (SR+DC) evaluations differ in their quality assurance (see Table 1). In both 2019 and 2020, references are treated as a "Human" system, to be ranked alongside the other systems; which may explain the lack of "REF" labeled segment types in the data. In 2019, the Appraise interface data used to generate the rankings did not include any segments labeled as "REPEAT", "REF", or "BAD_REF", though these are described as being included in the HITs (Barrault et al., 2019); perhaps they were removed before processing the data. In 2020, the Appraise data *did* include segments labeled as "BAD_REF", but none labeled as "REPEAT" or as "REF", while the 2020 Mechanical Turk document-level ones included all three. The 2019 data collected using the Turkle platform contains no human or reference data and we do not use it for any of our analysis in this work.

We reimplemented the scoring system using python and plan to release code for this paper. We were able to exactly replicate the raw scores and z-scores for most of the language pairs of interest from 2018-2020,[7] as well as the significance clusters.[8] See Appendix A for details. We use this reimplementation of the WMT scoring scripts in or-

[3] There is also a Document Rating with Document Context DR+DC, but we do not examine that in this work.

[4] A horizontal line is drawn below a system if and only if it is significantly better ($p < 0.05$) than *every* system with a lower z-score than it.

[5] We compute mean and standard deviation using `ad-latest.csv`, but use `ad-good-raw-redup.csv` to compute the individual z-scores and averages. The files are downloaded from 2018-2020 WMT websites: `http://www.statmt.org/wmt18/results.html`, `http://www.statmt.org/wmt19/results.html`, and `http://www.statmt.org/wmt20/results.html`.

[6] "SYSTEM" type are system outputs, while the remainder are quality assurance: "REPEAT" are repeated system outputs which are also valid for computing averages, "BAD_REF" are degraded references, and "REF" are references.

[7] In order to match the z-scores generated by the R packages used for WMT, we set `ddof` equal to 1 when using the `stats.zscore` function from scipy.

[8] We replicated the significance clusters using scipy's `stats.mannwhitneyu` function.

Dataset	SYSTEM	REPEAT	REF	BAD_REF
newstest2018-humaneval	265387	26489	26003	36924
appraise-doclevel-humaneval-newstest2019	194625	0	0	0
mturk-sntlevel-humaneval-newstest2019	92164	13266	13177	13113
turkle-sntlevel-humaneval-newstest2019	47799	0	0	0
appraise-doclevel-humaneval-newstest2020	186663	0	0	26856
mturk-sntlevel-humaneval-newstest2020	26262	3741	3746	3773
mturk-doclevel-humaneval-newstest2020	93777	12887	12939	12965

Table 1: Counts of sentence types in ad-good-raw-redup.csv files from 2018-2020. We omit the Turkle data from most of our analysis because it contains neither human systems nor reference data.

der to score authentic and modified WMT data, to examine underlying assumptions and hypothesize about how these may impact final system rankings.

For 21 language pairs annotated in SR–DC style and 25 in SR+DC style from 2018-2020, we were able to exactly replicate rankings, nearly replicate rankings (e.g., with rounding difference related changes to one significance line), or produce rankings whose differences could be explained by delays in data collection (2020 en-iu).[9] Appendix A provides more details on replication. We use our recalculated rankings and clusters as the starting point for all remaining analysis in this paper.

4 Understanding z-scores

While we've described how the z-score is calculated in the setting of the WMT human annotations, it's important to take a closer look at z-scores to understand how they behave in different scenarios. In this section, we explore z-scores and their underlying assumptions in hypothetical scenarios.

Given a raw score x, a mean μ, and a standard deviation σ, the **z-score** (or standard score) is the number of standard deviations above or below the mean that x falls. The z-score for a given raw score x can be computed as follows:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

This is a linear transformation; the shape of the distribution of z-scores is the same as that of the raw scores, but now with a mean of 0 and a standard deviation of 1. It is a unitless score.

Intuitively, the z-score provides a potential way of comparing scores from different annotators, but it requires a careful examination of underlying assumptions. If we think of the z-score as a unitless score, perhaps we can think of each annotator as having their own measurement units: we might have a lenient annotator and a harsh annotator, such that a raw score of 50 by the lenient annotator is quite *bad* while a raw score of 50 for the harsh annotator is actually quite *good*. In order to directly compare the two annotators' scores, we would like to map them to a shared scale, a unitless z-score. Under what assumptions is it appropriate to calculate z-scores to compare annotators' scores?

We start with perhaps the most obvious (but frequently unstated) assumption: there exists some latent "quality" of a given translation, which can be judged by a human annotator, such that annotators roughly agree about what constitutes a "good" or a "bad" translation. In practice, human annotators may disagree – for any number of reasons (Basile et al., 2021) – about which of two translations of "similar quality" is better, but we assume that the disagreement is not *extreme*; i.e., we hope that under a correlation coefficient like Pearson's r or Spearman's ρ, the correlation between annotators' scores would be much closer to 1 than to -1. For the sake of simplicity in the following examples, we will assume there exists a "true" and "objective" score for every translation.

Suppose that we have some translations with a true mean score of μ and a true standard deviation of σ. A lenient annotator scores all of the translations such that the distribution of their scores has a mean of $\mu + n$ and a standard deviation of σ, while a harsh annotator scores all of the translations such that the distribution of *their* scores has a mean of $\mu - m$ and a standard deviation of σ.[10] When we compute their z-scores, it is easier to directly compare sentence scores, since they are now on the same scale. This seems like a reasonable use of z-scores, but in this scenario annotators are scoring exactly the same data, which doesn't scale to WMT-style annotations; annotators simply don't

[9] Language codes: Chinese (zh), Czech (cs), German (de), English (en), Estonian (et), Finnish (fi), Gujarati (gu), Inuktitut (iu), Japanese (ja), Kazakh (kk), Khmer (km), Lithuanian (lt), Pashto (ps), Polish (pl), Russian (ru), Tamil (ta), Turkish (tr).

[10] We use the same standard deviation for simplicity, with arbitrary positive values of n and m.

have time to score all of the data.

Now suppose that we have two disjoint sets of sentences scored by two different annotators: the set S_X of sentences scored by annotator X and the set S_Y of sentences scored by annotator Y. From these raw scores, we can compute μ_X and μ_Y along with σ_X and σ_Y. If $\mu_X > \mu_Y$, can we conclude that annotator X is a more lenient annotator than annotator Y and resolve this by computing z-scores? Not without additional information! Imagine that we could see the "true" mean scores of S_X and S_Y, as annotated by a perfect omniscient annotator. It could be the case that the true means are identical and annotator X is indeed more lenient, but it could also be the case that the true mean of the scores in set S_X is actually higher. In the latter case, the annotators could be equally lenient, or it is even possible that annotator Y could be more lenient! In short, without a shared basis for comparison, we don't know whether computing z-scores is normalizing out annotator differences, differences in the data itself, or a combination.

5 z-scores in practice

This raises the question: what is happening in practice when we compute z-scores on WMT DAs? Are we really normalizing away inter-annotator differences, or is the normalization also doing something else, such as normalizing away real differences in HIT and system quality? If it is the latter, even z-scores for DAs may suffer the same bias from comparisons to better (or worse) systems.

We don't have access to an oracle, and we don't have a direct or reliable way to compute interannotator agreement, because in some collections it is rare that two annotators annotate the same text (and for the Appraise data, we only have HIT information, not annotator information). However, we can still examine this in the existing data and modifications thereof. Bojar et al. (2011) noted that systems might suffer from being compared to the reference too frequently under relative ranking, or might benefit from being compared to particularly poor systems. The same could hold true in DA. Consider the following toy example: a HIT contains 4 sentences, with raw scores of 25, 50, 50, 75, respectively. A sentence with a raw score of 50 in this HIT would have a z-score of 0. If, instead, the raw scores were 0, 25, 50, 75, a sentence with a raw score of 50 would have a z-score of 0.39, while for a HIT with raw scores of 25, 50, 75, 100, a sentence with a raw score of 50 would have a z-score of -0.39. While it is possible that such a set of scores could reflect differences in *annotator* behavior, we could also easily imagine that they might reflect differences in *HIT composition*, with one containing only system scores, one containing system scores and a bad reference, and one containing system scores and a (good) reference.

5.1 HIT Composition

Thus we examine HIT composition, or, more accurately, the composition of data annotated by any given worker/worker ID. In 2018, all systems were SR–DC, and 100% of workers annotated "BAD_REF" data.[11] However, an "Alternate DA HIT Structure" was employed for a subset of researcher HITs (run in Appraise), which used only "BAD_REF" segments for quality assurance, "omitting repeat pairs and good reference pairs" while also attempting to include "the output of all participating systems for each source input" (to have the same annotator produce annotations across systems). The percentage of (non-rejected) workers who annotated data containing "REF" in 2018 ranged from 4.9% (en-et) to 98.8% (zh-en); the former is an outlier, as the next two lowest are 25.8% (en-cs) and 47.4% (en-fi).

In 2019 annotations into English, 100% of workers annotated both "REF" and "BAD_REF" segments. In 2019 annotations out of English, the final output data does not include any "REF" or "BAD_REF" segments (though these *are* described as having been included for QA), but human references are treated as systems, and between 37.8% (en-de) and 61.5% (en-kk) of workers annotated at least some human reference data.

The 2020 Appraise annotations differed from prior years as well: 100% of the 2020 into English (Mechanical Turk) workers annotated both "REF" and "BAD_REF" segments. In 2020 annotations out of English (Appraise), between 95.8% (en-iu) and 100% (en-{ja, ta, zh}) of workers[12] annotated "BAD_REF" data. The percentage of Appraise "workers" that annotated data containing human references (treated as a system) ranged from 8.3% (en-iu) to 73.4% (en-zh).

[11] These values are calculated on ad-good-raw-redup.csv files, so only include annotators who successfully passed QA.

[12] The definition of "worker" is really a bit fuzzy here; the "WorkerID" produced by Appraise is really a HIT ID, so averages are *not* necessarily being computed across all of a given worker's annotations, but rather each HIT is being treated as a unique worker.

5.2 Analysis

In an ideal world where z-score normalization is only correcting for annotator variation, removing one system should not result in changes to the relative rankings of the remaining systems. That is to say, the z-scores themselves may be expected to change (shifting up if a very good system is removed, shifting down if a low-quality system is removed), but we wouldn't expect the relative ranking of two systems to change. After all, one stated motivation of the shift to DA was to avoid the known bias in RR of systems being unfairly penalized or benefiting unfairly from comparisons to stronger/weaker systems (Bojar et al., 2016). Similarly, replacing one system – for example with a much better or much worse system – should not result in other systems switching places in the rankings. We simulate these two scenarios using the existing data, and show that rankings produced in SR+DC settings are much more sensitive to removal or modification of systems than SR–DC.

Year/Type	Δ Rank	Δ Cluster	Δ Both
'18 (–DC)	1/13	0/13	0/13
'19 (–DC, MTurk)	2/5	1/5	0/5
'20 (–DC, MT.)	1/3	0/3	0/3
ALL SR–DC	**4/21**	**1/21**	**0/21**
'19 (+DC, MT.)	1/2	1/2	1/2
'19 (+DC, A.)	6/8	3/8	1/8
'20 (+DC, MT.)	7/7	3/7	3/7
'20 (+DC, A.)	4/8	5/8	3/8
ALL SR+DC	**18/25**	**12/25**	**8/25**

Table 2: Effect of removing human and "REF" scores from annotations and recalculating rankings by year, platform (MTurk or Appraise), and annotation style. Values indicate the fraction of language pairs that had changes in rank, clustering, or both rank and clustering.

We first examine removing human systems and "REF" – acting as though they had never been annotated at all, so all z-scores are calculated without "REF" or human system scores.[13] We then compute rankings and significance clusters. We compare these against the original rankings generated from all available data, with the significance clusters recomputed after removal of human systems.[14] For each pair of rankings, we check whether there is any change in the order of systems (ignoring significance clusters; we call this Δ *Rank*), whether there is any change in clusters (different number or composition of clusters; we call this Δ *Cluster*), and/or changes in both (Δ *Both*). Table 2 shows the results. Rank changes (ignoring significance clusters) are the most common, and many of these occur within significance clusters as we would expect. However, there are also a number of changes to the significance clusters (clusters merging, splitting, or rearranging), as well as pairs for which both rank and cluster changes occur. Most strikingly, all of these changes are *much* more common in the SR+DC settings than in the SR–DC. Removing human and "REF" data results in cluster changes to almost *half* (12/25) of the SR+DC rankings, but less than 5% (1/21) of the SR–DC rankings. No SR–DC rankings exhibit changes in both rank and clusters, but 32% of SR+DC rankings do. This is evidence that the SR+DC rankings are less stable, and consequently less reliable, than the SR–DC rankings. We replicate this result with removing the highest and lowest ranked systems, respectively, as shown in Table 3; the SR+DC rankings are much less robust than the SR–DC rankings to the removal of the best or worst single system.

Removed/Type	Δ Rank	Δ Cluster	Δ Both
Lowest (SR–DC)	4/21	3/21	1/21
Lowest (SR+DC)	18/25	10/25	7/25
Highest (SR–DC)	0/21	1/21	0/21
Highest (SR+DC)	17/25	10/25	5/25

Table 3: Effect of removing single lowest ranked or highest ranked system across all years, by data collection type (–/+DC). Values indicate the fraction of language pairs that had changes in rank, clustering, or both rank and clustering.

One might worry that some of this instability is due to the shrinking number of datapoints available when we remove "REF" and human systems, or the highest/lowest ranked systems. To account for this, we run the same experiment and measure the same changes, but instead of *removing* "REF" and human systems, we degrade their raw scores (dividing each score by 1.25, 1.5, 2, 4, and 10) before computing z-scores, rankings, and significance clusters. This could be viewed as a simulation of what would occur if the high-quality human system were replaced with mediocre (or, in the case of division by 10, very low-quality) systems.[15]

We visualize the result in Figure 1. Once again, the SR+DC evaluations are more brittle to these

[13] We observe similar results if we only remove "REF", but in that setting we cannot examine the 2019 and 2020 Appraise SR+DC rankings, as they do not make use of "REF" at all.

[14] Relevant to clusters containing or above human system(s).

[15] The reverse – inflating scores of low-performing systems – has a similar effect, but requires consideration of how to handle scores of zero.

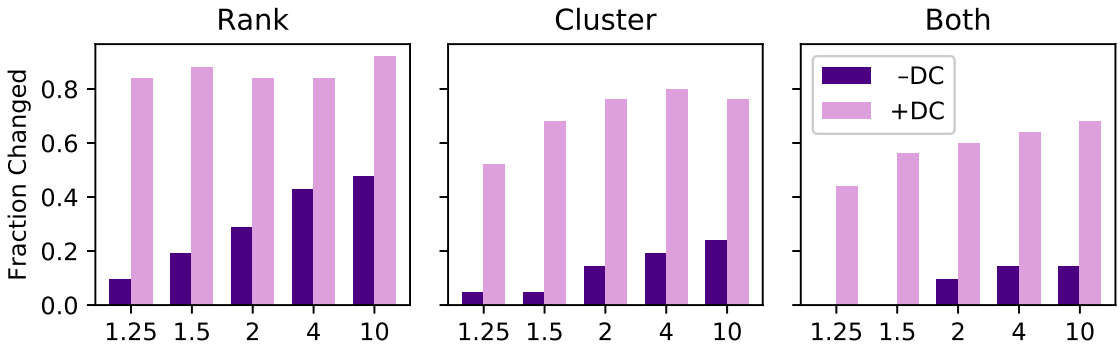

Figure 1: Effect of dividing raw human system and "REF" scores on overall (z-score) rankings for all SR–DC and SR+DC shown in Table 2. The x-axis shows the divisor (ranging from 1.25 to 10) and the y-axis shows the fraction of pairs for which the rankings, clusters, or both ranks and clusters changed.

changes. However, we see that even the SR–DC evaluations are not immune to the effects of extreme outliers on rankings and clusterings – as the divisor used increases, so does the fraction of pairs that have ranking and/or clustering changes. This makes sense intuitively: if most systems are of similar quality, a slight imbalance in which systems are compared to one another likely won't have dramatic effects, but if one system is much worse (or much better) than the rest, systems that are compared against it more or less frequently than others will see their z-scores benefit or suffer accordingly. We also examined monolingual vs. bilingual tasks in the SR+DC context (all SR–DC tasks were monolingual), but note similar rates of changes to ranks, clusters, and both across the two settings.

We have used a very coarse measurement here: counting whether the ranks or clusters changed *at all* rather than whether multiple clusters or large numbers of systems were reranked. Indeed, many of these changes are quite subtle, with just a single new significance line appearing or two clusters merging, or two close systems switching ranks (within or across clusters). If that is the case, why should we be concerned with this? The first reason is to better understand what it is that is actually being measured and whether the WMT annotation protocol is succeeding in its goals. If the inclusion of outliers or the degradation of system scores results in other systems shifting ranks, this indicates that the current approach does suffer from a similar comparison bias to RR. Thus we can't always be confident that what is being measured is a property of the system itself and not closely intertwined with HIT composition – this approach is doing something other than *only* normalizing away interannotator differences. The second reason is to highlight these goals and assumptions so that they can be considered when making future modifications to the annotation process. Many of these issues are currently resulting in small inconsistencies, but if future modifications are made to the annotation process without considering the underlying assumptions and goals, there is no reason to expect that the errors will cancel one another out rather than compound. If we are aware of the underlying assumptions when changes are introduced to the annotation process, we will be better positioned to consider potential problems in the hypothetical and then examine the real data to see if they appear in practice. There is also the question of effects on downstream tasks (§7). Finally, it also helps us to consider ways to mitigate these challenges before they grow, and we discuss some options for future consideration in §8.

5.3 Case Study

We manually select for examination a relatively dramatic case of rankings and clusters changing, from en-de 2020, pictured in Figure 2. This is an unusual case since it contained multiple human-based systems.[16]Nevertheless, it incorporates several issues we raised in hypotheticals, so we discuss it here.

Figure 2 shows the rankings for the original data (human systems were dropped only for the purpose of computing clusters, but *were* used for calculating z-scores), and each of the rankings computed by degrading raw scores by dividing them by 1.25 through 10 (denoted d-n where n is the divisor). We begin by focusing on PROMT_NMT, whose rank increases with increased degradation

[16]See Appendix A for details.

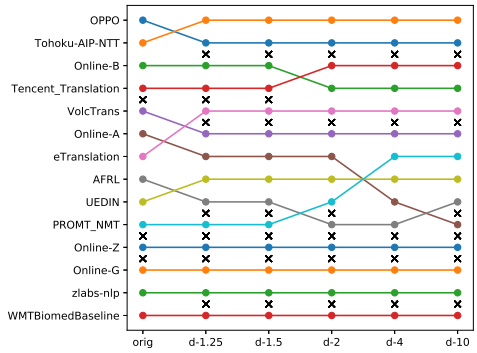

Figure 2: Plot showing z-score rankings (top is best) for 2020 en-de (SR+DC), from original rankings and five divisors for raw human scores. Significance lines are marked with black "x". Human systems were used in calculating z-scores but were removed prior to computing clusters for ease of visualization and comparison.

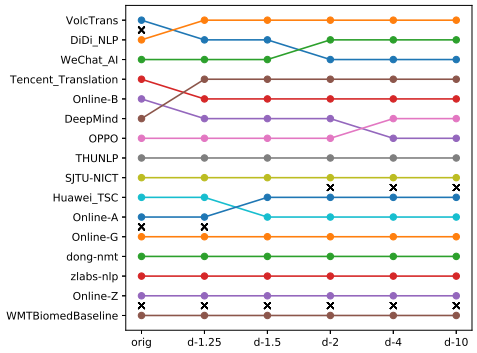

Figure 3: Rankings for 2020 zh-en (SR+DC), from original rankings and with divisors, as in Figure 2.

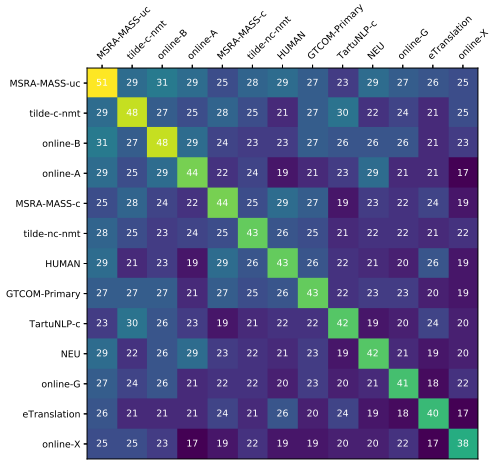

Figure 4: Co-occurrence matrix of systems for en-lt 2019 (SR+DC). Each cell shows the number of HITs that contained segments from the systems at those x and y values. The diagonal shows the total number of HITs that contained each system.

of the human systems. In the original ranking, AFRL and PROMT_NMT appear in the same cluster, with AFRL having a higher score than PROMT_NMT, but not statistically significantly so. When degrading the human raw scores by 1.25 or 1.5, AFRL is in a higher significance cluster than PROMT_NMT, but when dividing by 2, this is reversed: PROMT_NMT is now ranked as significantly *better* than AFRL, while with a divisor of 4 or 10, they return to the same cluster but with PROMT_NMT scoring higher. Thus we see, purely by degrading the raw scores of *other* systems, we observe the full range of possible relative rankings and clusterings for this pair of systems. The same holds true for PROMT_NMT compared with Online-A.

The en-de 2020 rankings may have suffered somewhat from having fewer annotations (1123.6 assessments per system), so we also show results for one of the most-assessed pairs that year: zh-en (2035.1 assessments per system). This is shown in Figure 3.[17] Here we focus on the top system: VolcTrans, which was ranked in Barrault et al. (2020) as significantly better than all systems. As we degrade the human systems, we see it begin to drop in rank, and this significance cluster merges with the one below it, raising the possibility that the initial finding was an artifact of the distribution of data across HITs rather than an inherent property of the MT quality of that particular system.

[17]Note that in the original rankings shown, the human system was omitted when computing significance clusters, and in this case a new significance line (separating Online-A and Online-G appears) where it had been, which was not there in the published rankings that do include human systems.

5.4 System Comparisons

There is a distinct difference in the way that systems are distributed across HITs in the SR–DC and SR+DC annotation styles. In SR–DC, almost all HITs contain segments from every single system (though there is no guarantee that they appear in exactly equal proportions to one another).

In SR+DC, this is not the case, owing to the fact that HITs are limited to 100 segments, there are often 10 or more systems, and documents are often longer than 10 segments. This means that it may be numerically impossible for a given HIT to cover all systems. We see this in Figure 4. A given system may be paired with any other system in less than half of the HITs in which it appears. These kinds

of imbalances mean that systems may be more frequently compared to better or worse systems, resulting in unfair effects on their rankings.

6 Documents

In both SR+DC and SR–DC styles, we don't have a guarantee that every segment-system pair is judged by an annotator, nor that at least one segment from every document-system pair is judged. If we assume approximately uniform translation difficulty across the test set, this isn't necessarily too much of a concern. However, is that really the case, or are some documents "easy" and others "hard"?

Figure 5 shows a matrix of document-system pairs, with each cell showing the average of all of the segment raw scores for that system-document pair.[18] The documents are ranked from highest average raw score to lowest average score (top to bottom), while the systems are ranked by highest average raw score to lowest average raw score (left to right). In the leftmost column, we see the "HUMAN" system, which has high scores across all documents. If all documents were equally difficult to translate, we would expect to see a gradient along the x-axis (i.e., across systems), with minimal variation along the y-axis (i.e., across documents). What we observe instead in this en-lt pair from 2019 (and across a number of other language pairs) is a rough gradient from the top left to the bottom right (with the exception of the "HUMAN" system, which remains strong throughout). This suggests that there are some documents that are "easy" for most systems to translate (top) and some that are "hard" (bottom). This raises a concern: when we attempt to compare two systems of very similar quality, they are not being measured on the same test set. An unlucky sample of documents might see one system judged on a "harder" set of documents, calling the resulting rankings into question.

7 Downstream Consequences

While researchers building MT systems for the shared task may view the human judgment rankings as the end result, the rankings are the *input* to the metrics tasks at WMT. Thus the reliability of the rankings has a direct impact on the reliability of the metrics task – which in the long term feeds into MT research as researchers decide

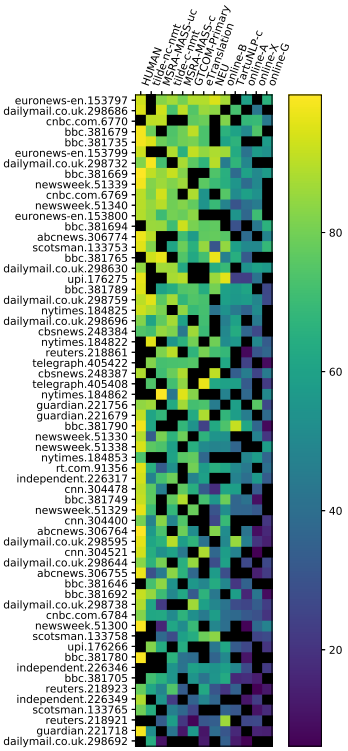

Figure 5: Average raw scores for document-system pairs from en-lt 2019 (SR+DC). Empty cells indicate pair was not judged. Documents are ranked by average raw score (highest: top) as are systems (highest: left).

which automatic metrics to use for evaluating their systems. In system-level metric evaluation at the WMT Metrics shared task, Pearson correlations are computed between metric scores and the z-score human rankings (Mathur et al., 2020b). Note that these correlations are directly between the system average z-scores and the metric scores, and as such do *not* treat all systems within a given cluster as tied. In practice, this means that even rank-only perturbations in the official ranking can be expected to cause changes to metrics task results.

Metrics scores are run on the *full* test set, not the various human-annotated subsets. Citing Graham et al. (2013), the Metrics task papers note that system-level DA scores are "consistent and have been found to be reproducible" even though different sets of segments are assessed for each system. However, that work predates the shift to sampling by document, and our analysis of instability and document difficulty suggest revisiting it.

Recent work has shown that outliers have a concerning impact on metric correlations (Mathur et al., 2020a), and organizers have worked to mitigate this (Mathur et al., 2020b). This paper is a step

[18] We can also produce such a matrix using z-scores or automatic metric scores, and results are comparable.

towards answering questions raised in Mathur et al. (2020b) regarding outliers and unfair advantages. It may seem tempting to remove outliers from human judgment tasks, but this will not solve the other problems and could instead mask their presence.

8 Proposals for Future Work

The issues discussed in this paper raise concerns about changes to the human evaluation protocols used at WMT and their effects on the validity of WMT system rankings. A partial solution would be to return to SR–DC annotations, perhaps after validation of the 2018 alternate HIT structure that guarantees that for every segment in the HIT, the HIT contains *every* MT system's output for that sentence. But this may be an unsatisfactory conclusion, and fails to address the interest in pushing MT evaluation toward whole documents.

Document-level and context-inclusive evaluations are growing in popularity, but there is limited study on document-level assessment methodologies for MT. Castilho (2021) examines setups comparable to SR–DC, SR+DC, and document rating with document context (which we omitted from this work), and finds in a controlled experiment using Likert scale ratings that a methodology comparable to SR+DC produces higher levels of interannotator agreement and fewer misevaluations than either whole document scores or individual sentences without context. However, that experimental setup does not suffer from the same task composition issues we observe in WMT; in fact these may be orthogonal issues.

If the choice is made to use SR+DC style annotations, there are some improvements to consider, but as noted in Castilho (2020), it remains "essential to test which methodologies will be best suited for different tasks and domains" prior to adopting them. One option would be to create 2018-alternate-structure style HITs with document context, where a HIT contains all systems' output for one or more documents. The downside to this is that it would likely require longer HITs or HITs that only contain a small number of documents; if systems are of similar quality, we might be concerned about annotator fatigue from repetition. The amount of context needed to adequately assess translations is still a question under consideration (Castilho et al., 2020; Castilho, 2021), which ties into issues of document and HIT length.

Another possibility to consider would be to always normalize over annotators (rather than over HITs), but this isn't a solution on its own – it is still necessary to make sure that annotators see comparable distributions of systems and documents, or the same problems will be reintroduced. Having annotators do calibration HITs, i.e., a set of annotations that *all* annotators complete, could also be considered. The calibration HITs would provide a consistent basis for computing the parameters of an annotator-specific z-score transformation, which could then be applied to the remainder of the annotator's judgments. This could untangle the issue of annotator strictness/leniency, but would still merit study before implementation (as annotator behavior may depend on HIT composition, so the z-scores learned in calibration may not be as applicable as one might hope if there is a mismatch between calibration HITs and the remainder of the HITs). One could also consider additional ways of modeling annotator behavior beyond z-score normalization (Paun et al., 2018).

A simpler starting point to deal with the issue of different systems being annotated over different documents would be to guarantee that all systems are scored over the same subset of documents.

All of these are (partially) orthogonal to the questions of what type of annotation tasks result in the most reliable ratings – whether it be direct assessment, ranking, or detailed error annotation – or questions of annotator skills and knowledge (Freitag et al., 2021).

9 Conclusions

We have shown that the current judgment collection methodology at the WMT news translation task results in SR+DC judgments that are more prone to variation on the basis of outliers than SR–DC judgments, and that HIT composition issues have helped reintroduce the relative ranking problem of unfair comparisons to the WMT rankings. We examined issues of document difficulty and how this interacts with the decision to sample documents (rather than sentences) for judgment. These issues risk undermining the validity of WMT rankings, with real consequences for MT research and downstream tasks on automatic metrics. In examining these issues, we've also presented several approaches to diving into the WMT ranking data that may be helpful to consider when planning future changes to WMT human judgment collection procedures.

Acknowledgments

Thank you to the anonymous reviewers for their suggestions and comments. Thank you to Chi-kiu Lo, Nitika Mathur, Gabriel Bernier-Colborne, Roland Kuhn, Huda Khayrallah, Adam Poliak, and George Foster for feedback on various drafts of this work. Thank you to Yvette Graham and task organizers for the 2020 data release and pointers to DA-related code. Thank you also to those listed above and a number of other current and former colleagues – including Rachel Rudinger, Eric Joanis, Darlene Stewart, Samuel Larkin, Michel Simard, Serge Léger, Patrick Littell, and Cyril Goutte – for discussions on related topics.

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

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.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

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*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, 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: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A grain of salt for the WMT manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 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. Association for Computational Linguistics.

Sheila Castilho. 2020. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online. Association for Computational Linguistics.

Sheila Castilho. 2021. Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45, Online. Association for Computational Linguistics.

Sheila Castilho, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France. European Language Resources Association.

Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *CoRR*, abs/2104.14478.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill™: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

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. Association for Computational Linguistics.

A Notes on Replication

As shown in Table 4, we are able to duplicate the following rankings exactly (or with minor differences, as noted). Code to replicate this work will be available at https://github.com/nrc-cnrc/WMT-Stability/. Language codes are as follows: Chinese (zh), Czech (cs), German (de), English (en), Estonian (et), Finnish (fi), Gujarati (gu), Inuktitut (iu), Japanese (ja), Kazakh (kk), Khmer (km), Lithuanian (lt), Pashto (ps), Polish (pl), Russian (ru), Tamil (ta), Turkish (tr).

- 2018, Mechanical Turk, SR–DC: en-{cs, de, et, fi, ru, tr, zh} and {cs, de, et, fi, ru, zh}-en, but we do not successfully replicate the scores for tr-en (we omit tr-en 2018 from future experiments).
- 2019, Appraise, SR+DC: en-{cs, de, fi, gu, kk, lt, ru, zh}, though we note that en-kk contains a duplicate system that is omitted from the published table.
- 2019, Mechanical Turk, SR–DC: {gu, kk, lt, ru}-en, and fi-en is nearly replicated, but our replication of it is missing a significance line between two clusters due to a rounding difference when computing the significance value.
- 2019, Mechanical Turk, SR+DC: {de, zh}-en are successfully replicated.
- 2019, Turkle, SR–DC: de-cs, de-fr, fr-de, zh-en, are all successfully replicated but are not included in the analyses.
- 2020, Appraise, SR+DC: en-{cs, ja, ru, ta, zh}, are successfully replicated, while en-pl is missing one significance line due to rounding differences. The ranking for en-de has identical scores *except* for Human-A and Human-paraphrase. The original en-de ranking in Barrault et al. (2020) included Human-A, Human-B, and Human-paraphrase. The released en-de data only contained Human-A and Human-B, though Human-A was about twice as large as Human-B, suggesting that it may have incorporated the Human-paraphrase data. Finally, the ranking for en-iu is quite different, though we expect this is because of delays in data collection resulting in a mismatch between the reported scores in the findings paper and the released scores. The en-iu scores also contain an additional low-scoring system that was omitted from the published table.
- 2020, Mechanical Turk, SR+DC: {cs, de, ja, pl, ru, ta, zh}-en were all replicated exactly.
- 2020, Mechanical Turk, SR–DC: {iu, km, ps}-en were all replicated exactly.
- 2020 en-{km, ps} appear to be missing from the released data.

Lang.	Year	–DC	+DC	Mono./Bi.	Tool	Replicated/Notes
en-cs	18	✓		M		✓
en-de	18	✓		M		✓
en-et	18	✓		M		✓
en-fi	18	✓		M		✓
en-ru	18	✓		M		✓
en-tr	18	✓		M		✓
en-zh	18	✓		M		✓
cs-en	18	✓		M		✓
de-en	18	✓		M		✓Matches clusters from Table 15, Appendix A, not Table 8.
et-en	18	✓		M		✓
fi-en	18	✓		M		✓
ru-en	18	✓		M		✓
*tr-en	18	✓		M		*Not successfully replicated.*
zh-en	18	✓		M		✓
en-cs	19		✓	B	Appraise	✓
en-de	19		✓	B	Appraise	✓
en-fi	19		✓	B	Appraise	✓
en-gu	19		✓	B	Appraise	✓
en-kk	19		✓	B	Appraise	✓Contains a duplicate of one system.
en-lt	19		✓	B	Appraise	✓
en-ru	19		✓	B	Appraise	✓
en-zh	19		✓	B	Appraise	✓Matches clusters from Table 33, Appendix A, not Table 11.
de-en	19		✓	M	MTurk	✓
fi-en	19	✓		M	MTurk	Missing a significance line (rounding difference).
gu-en	19	✓		M	MTurk	✓
kk-en	19	✓		M	MTurk	✓
lt-en	19	✓		M	MTurk	✓
ru-en	19	✓		M	MTurk	✓Matches clusters from Table 45, Appendix A, not Table 11.
zh-en	19		✓	M	MTurk	✓Note that this appears in Table 15.
*de-cs	19	✓		M	Turkle	✓
*de-fr	19	✓		M	Turkle	✓
*fr-de	19	✓		M	Turkle	✓
*zh-en	19	✓		M	Turkle	✓This is the Table 11 ranking.
en-cs	20		✓	B	Appraise	✓
en-de	20		✓	B	Appraise	Human-paraphrase missing (subsumed under Human-A?).
en-iu	20		✓	B	Appraise	Different scores/different data? Contains additional system.
en-ja	20		✓	B	Appraise	✓
*en-km	20			?	?	*Does not appear to exist.*
en-pl	20		✓	B	Appraise	Missing a significance line (rounding difference).
*en-ps	20			?	?	*Does not appear to exist.*
en-ru	20		✓	B	Appraise	✓
en-ta	20		✓	B	Appraise	✓
en-zh	20		✓	B	Appraise	✓
cs-en	20		✓	M	MTurk	✓
de-en	20		✓	M	MTurk	✓
iu-en	20	✓		M	MTurk	✓
ja-en	20		✓	M	MTurk	✓
km-en	20	✓		M	MTurk	✓
pl-en	20		✓	M	MTurk	✓
ps-en	20	✓		M	MTurk	✓
ru-en	20		✓	M	MTurk	✓
ta-en	20		✓	M	MTurk	✓
zh-en	20		✓	M	MTurk	✓

Table 4: Notes on Findings paper ranking replications, including information about language pairs, year, SR–DC vs. SR+DC, monolingual vs. bilingual evaluation, tool used for data collection, and success or failure to replicate. Systems marked with * were not included in any additional analysis.