Experts, Errors, and Context:
A Large-Scale Study of Human Evaluation for Machine Translation

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, Wolfgang Macherey
Google Research
{freitag, fosterg, grangier, vratnakar, qijuntan, wmach}@google.com

Abstract

Human evaluation of modern high-quality machine translation systems is a difficult problem, and there is increasing evidence that inadequate evaluation procedures can lead to erroneous conclusions. While there has been considerable research on human evaluation, the field still lacks a commonly-accepted standard procedure. As a step toward this goal, we propose an evaluation methodology grounded in explicit error analysis, based on the Multidimensional Quality Metrics (MQM) framework. We carry out the largest MQM research study to date, scoring the outputs of top systems from the WMT 2020 shared task in two language pairs using annotations provided by professional translators with access to full document context. We analyze the resulting data extensively, finding among other results a substantially different ranking of evaluated systems from the one established by the WMT crowd workers, exhibiting a clear preference for human over machine output. Surprisingly, we also find that automatic metrics based on pre-trained embeddings can outperform human crowd workers. We make our corpus publicly available for further research.

1 Introduction

Like many natural language generation tasks, machine translation (MT) is difficult to evaluate because the set of correct answers for each input is large and usually unknown. This limits the accuracy of automatic metrics, and necessitates costly human evaluation to provide a reliable gold standard for measuring MT quality and progress. Yet even human evaluation is problematic. For instance, we often wish to decide which of two translations is better, and by how much, but what should this take into account? If one translation sounds somewhat more natural than another, but

contains a slight inaccuracy, what is the best way to quantify this? To what extent will different raters agree on their assessments?

The complexities of evaluating translations—both machine and human—have been extensively studied, and there are many recommended best practices. However, due to expedience, human evaluation of MT is frequently carried out on isolated sentences by inexperienced raters with the aim of assigning a single score or ranking. When MT quality is poor, this can provide a useful signal; but as quality improves, there is a risk that the signal will become lost in rater noise or bias. Recent papers have argued that poor human evaluation practices have led to misleading results, including erroneous claims that MT has achieved human parity (Toral, 2020; Läubli et al., 2018).

This paper aims to contribute to the evolution of standard practices for human evaluation of high-quality MT. Our key insight is that any scoring or ranking of translations is implicitly based on an identification of errors and other imperfections. Making such an identification explicit by enumerating errors provides a "platinum standard" from which various gold-standard scorings can be derived, depending on the importance placed on different categories of errors for different downstream tasks. This is not a new insight: it is the conceptual basis for the Multidimensional Quality Metrics (MQM) framework developed in the EU QTLaunchPad and QT21 projects (www.qt21.eu), which we endorse and adopt for our experiments.

MQM is a generic framework that provides a hierarchy of translation errors which can be tailored to specific applications. We identified a hierarchy appropriate for broad-coverage MT, and annotated outputs from 10 top-performing "systems" (including human references) for both the English→German (EnDe) and Chinese→English (ZhEn) language directions in the WMT 2020 news translation task (Barrault et al., 2020), using

professional translators with access to full document context. For comparison purposes, we also collected scalar ratings on a 7-point scale from both professionals and crowd workers.

We analyze the resulting data along many different dimensions: comparing the system rankings resulting from different rating methods, including the original WMT scores; characterizing the error patterns of modern neural MT systems, including profiles of difficulty across documents, and comparing them to human translations (HT); measuring MQM inter-annotator agreement; and re-evaluating the performance of automatic metrics submitted to the WMT 2020 metrics task. Our most striking finding is that MQM ratings sharply revise the original WMT ranking of translations, exhibiting a clear preference for HT over MT, and promoting some low-ranked MT systems to much higher positions. This in turn changes the conclusions about the relative performance of different automatic metrics; interestingly, we find that most metrics correlate better with MQM rankings than WMT human scores do. We hope these results will underscore and help publicize the need for more careful human evaluation, particularly in shared tasks intended to assess MT or metric performance. We release our corpus to encourage further research.[1] Our main contributions are:

- A proposal for a standard MQM scoring scheme appropriate for broad-coverage MT.

- Release of a large-scale MQM corpus with annotations for over 100k HT and high-quality-MT segments in two language pairs (EnDe and ZhEn) from WMT 2020. This is by far the largest study of human evaluation results released to the public.

- Re-evaluation of the performance of MT systems and automatic metrics on our corpus, showing clear distinctions between HT and MT based on MQM ratings, adding to the evidence against claims of human parity.

- Demonstration that crowd-worker evaluation has low correlation with our MQM-based evaluation, calling into question conclusions drawn on the basis of previous crowd-sourced evaluations.

---

[1] https://github.com/google/wmt-mqm-human-evaluation

- Demonstration that automatic metrics based on pre-trained embeddings can outperform human crowd workers.

- Characterization of current error types in HT and MT, identifying specific MT weaknesses.

- Recommendations for the number of ratings needed to establish a reliable human benchmark, and for the most efficient way of distributing them across documents.

2 Related Work

One of the earliest formal mentions of human evaluation for MT occurs in the ALPAC report (1966), which defines an evaluation methodology based on "intelligibility" (comprehensibility) and "fidelity" (adequacy). The ARPA MT Initiative (White et al., 1994) defines an overall quality score based on "adequacy", "fluency" and "comprehension". In 2006, the first WMT evaluation campaign (Koehn and Monz, 2006) used adequacy and fluency ratings on a 5 point scale acquired from participants as their main metric. Vilar et al. (2007) proposed a ranking-based evaluation approach which became the official metric at WMT from 2008 until 2016 (Callison-Burch et al., 2008). The ratings were still acquired from the participants of the evaluation campaign. Graham et al. (2013) compared human assessor consistency levels for judgments collected on a five-point interval-level scale to those collected on a 1-100 continuous scale, using machine translation fluency as a test case. They claim that the use of a continuous scale eliminates individual judge preferences, resulting in higher levels of inter-annotator consistency. Bojar et al. (2016) came to the conclusion that fluency evaluation is highly correlated to adequacy evaluation. As a consequence of the latter two papers, continuous direct assessment focusing on adequacy has been the official WMT metric since 2017 (Bojar et al., 2017). Due to budget constraints, WMT understandably conducts its human evaluation with researchers and/or crowd-workers.

Avramidis et al. (2012) used professional translators to rate MT output on three different tasks: ranking, error classification and post-editing. Castilho et al. (2017) found that crowd workers lack knowledge of translation and, compared to professional translators, tend to be more accepting of (subtle) translation errors. Graham et al.

(2017) showed that crowd-worker evaluation has to be filtered to avoid contamination of results through the inclusion of false assessments. The quality of ratings acquired by either researchers or crowd workers has further been questioned by (Toral et al., 2018; Läubli et al., 2020), who demonstrated that professional translators can discriminate between human and machine translations where crowd-workers were not able to do so. Mathur et al. (2020) re-evaluated a subset of WMT submissions with professional translators and showed that the resulting rankings changed and were better aligned with automatic scores. Fischer and Läubli (2020) found that the number of segments with wrong terminology, omissions, and typographical problems for MT output is similar to HT. Fomicheva et al. (2017); Bentivogli et al. (2018) raised the concern that reference-based human evaluation might penalise correct translations that diverge too much from the reference. The literature mostly agrees that source-based rather than reference-based evaluation should be conducted (Läubli et al., 2020). The impact of translationese (Koppel and Ordan, 2011) on human evaluation of MT has recently received attention (Toral et al., 2018; Zhang and Toral, 2019; Freitag et al., 2019; Graham et al., 2020). These papers show that the nature of source sentences is important and that only natural source sentences should be used for human evaluation.

As alternatives to adequacy and fluency, Scarton and Specia (2016) presented reading comprehension for MT quality evaluation. Forcada et al. (2018) proposed gap-filling, where certain words are removed from reference translations and readers are asked to fill the gaps left using the machine-translated text as a hint. Popović (2020) proposed a new method for manual evaluation based on marking actual issues in the translated text. Instead of assigning a score, annotators are asked to just label problematic parts of the translations.

The Multidimensional Quality Metrics (MQM) framework was developed in the EU QT-LaunchPad and QT21 projects (2012–2018) (www.qt21.eu) to address the shortcomings of previous quality evaluation methods (Lommel et al., 2014). MQM provides a generic methodology for assessing translation quality that can be adapted to a wide range of evaluation needs. Klubička et al. (2018) designed an MQM-compliant error taxonomy tailored to the relevant linguistic phe-nomena of Slavic languages to run a case study for 3 MT systems for English→Croatian. More recently, Rei et al. (2020) used MQM labels to fine-tune COMET for automatic evaluation.

3 Human Evaluation Methodologies

We compared three human evaluation techniques: the WMT 2020 baseline; ratings on a 7-point Likert-type scale which we refer to as a Scalar Quality Metric (SQM); and evaluations under the MQM framework. We describe these methodologies in the following three sections, deferring concrete experimental details about annotators and data to the subsequent section.

3.1 WMT

As part of the WMT evaluation campaign (Barrault et al., 2020), WMT runs human evaluation of the primary submissions for each language pair. The organizers collect segment-level ratings with document context (SR+DC) on a 0-100 scale using either source-based evaluation with a mix of researchers/translators (for translations out of English) or reference-based evaluation with crowd-workers (for translations into English). In addition, WMT conducts rater quality controls to remove ratings from raters that are not trustworthy. In general, for each system, only a subset of documents receive ratings, with the rated subset differing across systems. The organizers provide two different segment-level scores, averaged across one or more raters: (a) the raw score; and (b) a z-score which is standardized for each annotator. Document- and system-level scores are averages over segment-level scores. For more details, we refer the reader to the WMT findings papers.

3.2 SQM

Similar to the WMT setting, the Scalar Quality Metric (SQM) evaluation collects segment-level scalar ratings with document context. Different from the 0-100 assessment of translation quality used in WMT, SQM uses a 0-6 scale for translation quality assessment, with the quality levels described as follows:

6: Perfect Meaning and Grammar: The meaning of the translation is completely consistent with the source and the surrounding context (if applicable). The grammar is also correct.

4: Most Meaning Preserved and Few Grammar Mistakes: The translation retains most of the

meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.

2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.

0: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source. Grammar is irrelevant.

This evaluation presents each source segment and translated segment from a document in a table row, asking the rater to pick a rating from 0 through 6 (including the intermediate levels 1, 3, and 5). The rater can scroll up or down to see all the other source/translation segments from the document. Our SQM experiments used the 0-6 rating scale described above, instead of the wider, continuous scale recommended by (Graham et al., 2013), as this scale has been an established part of our existing MT evaluation ecosystem. It is possible that system rankings may be slightly sensitive to this nuance, but less so with raters who are translators rather than crowd workers, we believe.

3.3 MQM

To adapt the generic MQM framework for our context, we followed the official guidelines for scientific research (MQM-usage-guidelines.pdf). For space reasons we give only the salient features of our MQM customization here, referring the reader to appendix A for a summary of MQM, and to appendix B for full details of our framework.

Our annotators were instructed to identify all errors within each segment in a document, paying particular attention to document context; see Table 12 for complete annotator guidelines. Each error was highlighted in the text, and labeled with an error category from Table 10 and a severity from Table 11. To temper the effect of long segments, we imposed a maximum of five errors per segment, instructing raters to choose the five most severe errors for segments containing more errors.

Our error hierarchy includes the standard top-level categories *Accuracy*, *Fluency*, *Terminology*, *Style*, and *Locale*, each with a specific set of subcategories. After an initial pilot run, we introduced a special *Non-translation* error that can be used to tag an entire segment which is too badly garbled to permit reliable identification of individual errors.

Error severities are assigned independent of cat-egory, and consist of *Major*, *Minor*, and *Neutral* levels, corresponding respectively to actual translation or grammatical errors, smaller imperfections, and purely subjective opinions about the translation. Many MQM schemes include an additional *Critical* severity which is worse than Major, but we dropped this because its definition is often context-specific. We felt that for broad coverage MT, the distinction between Major and Critical was likely to be highly subjective, while Major errors (true errors) would be easier to distinguish from Minor ones (imperfections).

Since we are ultimately interested in scoring segments, we require a weighting on error types. We fixed the weight on Minor errors at 1, and explored a range of Major weights from 1 to 10 (the Major weight recommended in the MQM standard). For each weight combination we examined the stability of system ranking using a resampling technique. We found that a Major weight of 5 gave the best balance between stability and ability to discriminate among systems.

These weights apply to all error categories with two exceptions. We assigned a weight of 0.1 to Minor Fluency/Punctuation errors to reflect their mostly non-linguistic nature. Decisions like the style of quotation mark to use or the spacing around punctuation affect the appearance of a text but do not change its meaning. Unlike other kinds of Minor errors, these are easy to correct algorithmically, so we assign a low weight to ensure that their main role is to distinguish between systems that are equivalent in other respects. Major Fluency/Punctuation errors, which render a text ungrammatical or change its meaning (eg, eliding the comma in "Let's eat, grandma"), have standard weighting. The second exception is the singleton Non-translation category, with a weight of 25, equivalent to five Major errors.

Table 1 summarizes our weighting scheme, in which segment-level scores can range from 0 (perfect) to 25 (worst). The final segment-level score is an average over scores from all annotators.

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	all	0

Table 1: MQM error weighting.

3.4 Experimental Setup

We re-annotated the WMT 2020 English→German and Chinese→English test sets, comprising 1418 segments (130 documents) and 2000 segments (155 documents) respectively. For each set we chose 10 "systems" for annotation, including the three reference translations available for English→German and the two references available for Chinese→English. The MT outputs included all top-performing systems according to the WMT human evaluation, augmented with systems we selected to increase diversity. Tables 3 and 4 list all evaluated systems.

Table 2 summarizes rating information for the WMT evaluation and for the additional evaluations we conducted: SQM with crowd workers (cSQM), SQM with professional translators (pSQM), and MQM. We used disjoint professional translator pools for pSQM and MQM in order to avoid bias. All members of our rater pools were native speakers of the target language. Note that the average number of ratings per segment is less than 1 for the WMT evaluations because not all ratings survived the quality control.

	ratings / seg	rater pool	raters
WMT EnDe	0.47	res./trans.	115
WMT ZhEn	0.86	crowd	219
cSQM EnDe	3	crowd	276
cSQM ZhEn	1	crowd	70
pSQM	3	professional	6
MQM	3	professional	6

Table 2: Details of all human evaluations.

To ensure maximum diversity in ratings for pSQM and MQM, we assigned documents in round-robin fashion to all 20 different sets of 3 raters from these pools. We chose an assignment order that roughly balanced the number of documents and segments per rater. Each rater was assigned a subset of documents, and annotated outputs from all 10 systems for those documents. Both documents and systems were anonymized and presented in a different random order to each rater. The number of segments per rater ranged from 6,830–7,220 for English→German and from 9,860–10,210 for Chinese→English.

4 Results

4.1 Overall System Rankings

For each human evaluation setup, we calculate a system-level score by averaging the segment-level scores for each system. Results are summarized in Table 3 (English→German) and Table 4 (Chinese→English). The system- and segment-level correlations to our platinum MQM ratings are shown in Figure 1 and 2 (English→German), and Figure 3 and 4 (Chinese→English). Segment-level correlations are calculated only for segments that were evaluated by WMT. For both language pairs, we observe similar patterns when looking at the results of the different human evaluations and come to the following findings:
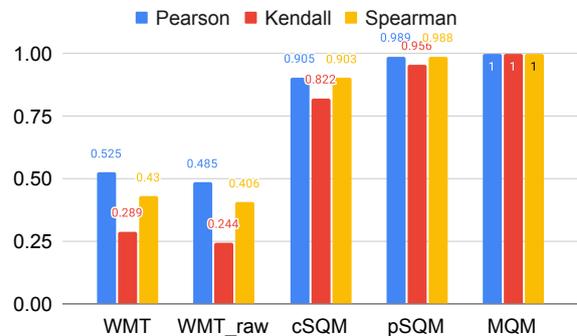

Figure 1: English→German: System correlation with the platinum ratings acquired with MQM.
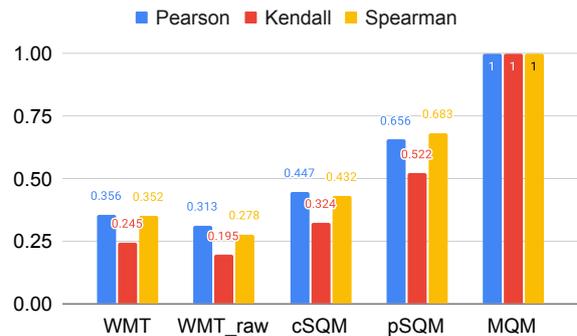

Figure 2: English→German: Segment-level correlation with the platinum ratings acquired with MQM.

(i) Human translations are underestimated by crowd workers: Already in 2016, Hassan et al. (2018) claimed human parity for news-translation for Chinese→English. We confirm the findings of Toral et al. (2018); Läubli et al. (2018) that when human evaluation is conducted correctly, professional translators can discriminate between human and machine translations. All human translations

5

System	WMT↑	WMT RAW↑	cSQM↑	pSQM↑	MQM↓	Major↓	Minor↓	Fluency↓	Accuracy↓
Human-B	0.569(1)	90.5(1)	5.31(1)	5.16(1)	0.75(1)	0.22(1)	0.54(1)	0.28(1)	0.47(1)
Human-A	0.446(4)	85.7(4)	5.20(2)	4.90(2)	0.91(2)	0.28(2)	0.64(2)	0.33(2)	0.58(2)
Human-P	0.299(10)	84.2(9)	5.04(5)	4.32(3)	1.41(3)	0.57(3)	0.85(3)	0.50(3)	0.91(3)
Tohoku-AIP-NTT	0.468(3)	88.6(2)	5.11(3)	3.95(4)	2.02(4)	0.94(4)	1.14(4)	0.61(5)	1.40(4)
OPPO	0.495(2)	87.4(3)	5.03(6)	3.79(5)	2.25(5)	1.07(5)	1.19(6)	0.62(6)	1.63(5)
eTranslation	0.312(9)	82.5(10)	5.02(7)	3.68(7)	2.33(6)	1.18(7)	1.16(5)	0.56(4)	1.78(7)
Tencent_Translation	0.386(6)	84.3(8)	5.06(4)	3.77(6)	2.35(7)	1.15(6)	1.22(8)	0.63(7)	1.73(6)
VolcTrans	0.326(7)	84.6(6)	5.00(8)	3.65(8)	2.45(8)	1.23(8)	1.23(9)	0.64(8)	1.80(8)
Online-B	0.416(5)	84.5(7)	4.95(9)	3.60(9)	2.48(9)	1.34(9)	1.20(7)	0.64(9)	1.84(9)
Online-A	0.322(8)	85.3(5)	4.85(10)	3.32(10)	2.99(10)	1.73(10)	1.32(10)	0.76(10)	2.23(10)

Table 3: English→German: Different human evaluations for 10 submissions of the WMT20 evaluation campaign.

System	WMT↑	WMT RAW↑	cSQM↑	pSQM↑	MQM↓	Major↓	Minor↓	Fluency↓	Accuracy↓
Human-A	-	-	5.09(2)	4.34(1)	3.43(1)	2.71(1)	0.74(1)	0.91(1)	2.52(1)
Human-B	-0.029(9)	74.8(9)	5.03(7)	4.29(2)	3.62(2)	2.81(2)	0.82(10)	0.95(2)	2.66(2)
VolcTrans	0.102(1)	77.47(5)	5.04(5)	4.03(3)	5.03(3)	4.26(3)	0.79(6)	1.31(7)	3.71(3)
WeChat_AI	0.077(3)	77.35(6)	4.99(8)	4.02(4)	5.13(4)	4.39(4)	0.76(4)	1.24(5)	3.89(4)
Tencent_Translation	0.063(4)	76.67(7)	5.04(6)	3.99(5)	5.19(5)	4.43(6)	0.79(8)	1.23(4)	3.96(5)
OPPO	0.051(7)	77.51(4)	5.07(4)	3.99(5)	5.20(6)	4.41(5)	0.81(9)	1.23(3)	3.97(6)
THUNLP	0.028(8)	76.48(8)	5.11(1)	3.98(7)	5.34(7)	4.61(7)	0.75(3)	1.27(6)	4.07(9)
DeepMind	0.051(6)	77.96(1)	5.07(3)	3.97(8)	5.41(8)	4.67(8)	0.75(2)	1.38(8)	4.02(7)
DiDi_NLP	0.089(2)	77.63(3)	4.91(9)	3.95(9)	5.48(9)	4.73(9)	0.77(5)	1.43(9)	4.05(8)
Online-B	0.06(5)	77.77(2)	4.83(10)	3.89(10)	5.85(10)	5.08(10)	0.79(7)	1.51(10)	4.34(10)

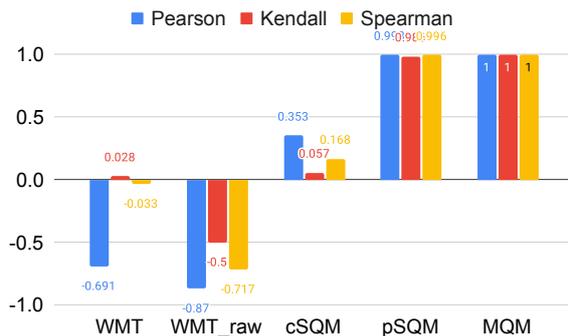Table 4: Chinese→English: Different human evaluations for 10 submissions of the WMT20 evaluation campaign.

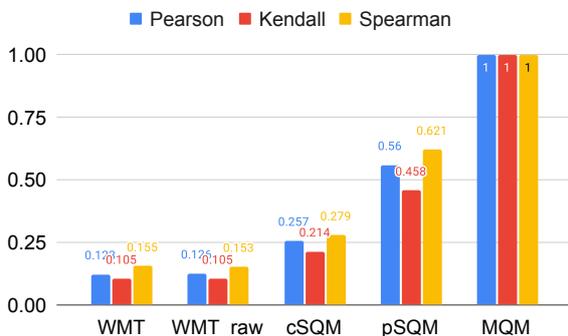Figure 3: Chinese→English: System-level correlation with the platinum ratings acquired with MQM.

Figure 4: Chinese→English: Segment-level correlation with the platinum ratings acquired with MQM.

are ranked first by both the pSQM and MQM evaluations for both language pairs. The gap between

human translations and MT is even more visible when looking at the MQM ratings which sets the human translations first by a large margin, demonstrating that the quality difference between MT and human translation is still large. Another interesting observation is the ranking of Human-P for English→German. Human-P is a reference translation generated using the paraphrasing method of (Freitag et al., 2020) which asked linguists to paraphrase existing reference translations as much as possible while also suggesting using synonyms and different sentence structures. Our results support the assumption that crowd-workers are biased to prefer literal, easy-to-rate translations and rank Human-P low. Professional translators on the other hand are able to see the correctness of the paraphrased translations and ranked them higher than any MT output. Similar to the standard human translations, the gap between Human-P and the MT systems is larger when looking at the MQM ratings. In MQM, raters have to justify their ratings by labelling the error spans which helps to avoid penalizing non-literal translations.

(ii) WMT has low correlation with MQM: The human evaluation in WMT was conducted by crowd-workers (Chinese→English) or a mix of researchers/translators (English→German) during the WMT evaluation campaign. Further, different to all other evaluations in this paper,

WMT conducted a reference-based/monolingual human evaluation for Chinese→English in which the machine translation output was compared to a human-generated reference. When comparing the system ranks based on WMT for both language pairs with the ones generated by MQM, we can see low correlation for English→German (see Figure 1) and even negative correlation for Chinese→English (see Figure 3). We also see very low segment-level correlation for both language pairs (see Figure 2 and Figure 4). Later, we will also show that the correlation of SOTA automatic metrics are higher than the human ratings generated by WMT. The results at least question the reliability of the human ratings acquired by WMT.

(iii) pSQM has high system-level correlation with MQM: The results for both language pairs suggest that pSQM and MQM are of similar quality as their system rankings mostly agree. Nevertheless, when zooming into the segment-level correlations, we observe a much lower correlation of ∼0.5 based on Kendall tau for both language pairs. The difference of the two approaches is also visible in the absolute differences of the individual systems. For instance the submissions of DiDi_NLP and Tencent_Translation for Chinese→English are close for pSQM (only 0.04 absolute difference). MQM on the other hand shows a larger difference of 0.19 points. When the quality of two systems gets closer, a more fine-grained evaluation schema like MQM is needed. This is also important when doing system development where the difference between two variations for two systems can be minor. Looking into the future when we get closer to human translation quality, MQM will be needed for reliable evaluation. On the other hand, pSQM seems to be sufficient for an evaluation campaign like WMT.

(iv) MQM results are mainly driven by major and accuracy errors: In Table 3 and Table 4, we also show the MQM error scores only based on Major/Minor errors or only based on Fluency or Accuracy errors. Interestingly, the MQM score based on accuracy errors or based on Major errors gives us almost the same rank as the full MQM score. Later in the paper, we will see that the majority of major errors are accuracy errors. This suggests the quality of an MT system is still driven mostly by accuracy errors as most fluency errors are judged minor.

4.2 Error Category Distribution

MQM provides fine-grained error categories grouped under 4 main categories (accuracy, fluency, terminology and style). The absolute error counts for all 3 ratings for all 10 systems are shown in Tables 5 and 6. The error category Accuracy/Mistranslation is responsible for the majority of major errors for both language pairs. This suggests that the main problem of MT is still mistranslation of words or phrases. The absolute number of errors is much higher for Chinese→English which demonstrates that this translation pair is more challenging than English→German.

Table 5 decomposes system and human MQM scores per category for English→German. Human translations get lower error counts in all categories, except for additions. It seems that human translators might add tokens for fluency which are not supported by the source. Both systems and humans are mostly penalized by accuracy/mistranslation errors, but systems record 4x more error points in these categories. Similarly, sentences with more than 5 major errors (nontranslation) are much more frequent for systems (∼ 28x the human rate). The best systems are quite different across categories. Tohoku is average in fluency but outstanding in accuracy, eTranslation is excellent in fluency but worse in accuracy, and OPPO ranks between the two other systems for both aspects. Compared to humans, the best systems are mostly penalized for mistranslations and non-translation (badly garbled sentences).

Table 6 shows that the Chinese→English translation task is more difficult than English→German translation, with higher MQM error scores for human translations. Again, humans are performing better than systems across all categories except for additions, omissions and spelling. Many spelling mistakes relate to name formatting and capitalization which is difficult for this language pair (see name formatting errors). Additions and omissions again highlight that humans might be ready to compromise accuracy for fluency in some cases. Mistranslation and name formatting are the categories where the systems are penalized the most compared to humans. When comparing systems, the differences between the best systems is less pronounced than for English→German, both in term of aggregate score and per-category counts.

Error Categories	Errors (%)	Major (%)	Human MQM	All MT MQM	All MT vs H.	Tohoku MQM	Tohoku vs H.	OPPO MQM	OPPO vs H.	eTrans MQM	eTrans vs H.
Accuracy/Mistranslation	33.2	27.2	0.296	1.285	*4.3*	1.026	*3.5*	1.219	*4.1*	1.244	*4.2*
Style/Awkward	14.6	4.6	0.146	0.299	*2.0*	0.289	*2.0*	0.315	*2.1*	0.296	*2.0*
Fluency/Grammar	10.7	4.7	0.097	0.224	*2.3*	0.193	*2.0*	0.215	*2.2*	0.196	*2.0*
Accuracy/Omission	3.6	13.4	0.070	0.091	*1.3*	0.063	*0.9*	0.063	*0.9*	0.120	*1.7*
Accuracy/Addition	1.8	6.7	0.067	0.025	*0.4*	0.018	*0.3*	0.024	*0.4*	0.021	*0.3*
Terminology/Inappropriate	8.3	7.0	0.061	0.193	*3.2*	0.171	*2.8*	0.189	*3.1*	0.193	*3.2*
Fluency/Spelling	2.3	1.2	0.030	0.039	*1.3*	0.030	*1.0*	0.039	*1.3*	0.028	*0.9*
Accuracy/Untranslated tex	3.1	14.9	0.024	0.090	*3.8*	0.082	*3.5*	0.066	*2.8*	0.098	*4.2*
Fluency/Punctuation	20.3	0.2	0.014	0.039	*2.8*	0.067	*4.9*	0.013	*1.0*	0.011	*0.8*
Other	0.5	5.2	0.005	0.010	*1.9*	0.009	*1.6*	0.010	*1.9*	0.007	*1.2*
Fluency/Register	0.6	5.0	0.005	0.014	*3.0*	0.009	*1.9*	0.015	*3.2*	0.015	*3.3*
Terminology/Inconsistent	0.3	0.0	0.004	0.005	*1.2*	0.004	*0.9*	0.005	*1.2*	0.005	*1.2*
Non-translation	0.2	100.0	0.003	0.083	*28.3*	0.041	*14.0*	0.065	*22.0*	0.094	*32.0*
Fluency/Inconsistency	0.1	1.3	0.003	0.002	*0.7*	0.001	*0.3*	0.001	*0.3*	0.003	*1.0*
Fluency/Character enc.	0.1	3.7	0.002	0.001	*0.7*	0.002	*1.0*	0.001	*0.6*	0.000	*0.2*
All accuracy	41.7	24.2	0.457	1.492	*3.3*	1.189	*2.6*	1.372	*3.0*	1.483	*3.2*
All fluency	34.2	1.8	0.150	0.320	*2.1*	0.303	*2.0*	0.284	*1.9*	0.253	*1.7*
All except acc. & fluenc	24.2	6.0	0.222	0.596	*2.7*	0.526	*2.4*	0.591	*2.7*	0.596	*2.7*
All categories	100.0	12.1	0.829	2.408	*2.9*	2.017	*2.4*	2.247	*2.7*	2.332	*2.8*

Table 5: Category breakdown of MQM scores for English→German for human translations (A, B), machine translations (all systems) and some of the best systems (Tohohku, OPPO, eTranslation). The ratio of system over human scores is in italics. Errors (%) report the fraction of the total error counts in a category, Major (%) report the fraction of major error for each category.

Error Categories	Errors (%)	Major (%)	Human MQM	All MT MQM	All MT vs H.	VolcTrans MQM	VolcTrans vs H.	WeChat MQM	WeChat vs H.	Tencent MQM	Tencent vs H.
Accuracy/Mistranslation	42.2	71.5	1.687	3.218	*1.9*	2.974	*1.8*	3.108	*1.8*	3.157	*1.9*
Accuracy/Omission	8.6	61.3	0.646	0.505	*0.8*	0.468	*0.7*	0.534	*0.8*	0.547	*0.8*
Fluency/Grammar	13.8	18.4	0.381	0.442	*1.2*	0.414	*1.1*	0.392	*1.0*	0.425	*1.1*
Locale/Name format	6.4	74.5	0.250	0.505	*2.0*	0.506	*2.0*	0.491	*2.0*	0.433	*1.7*
Terminology/Inappropriate	5.1	31.1	0.139	0.221	*1.6*	0.220	*1.6*	0.217	*1.6*	0.202	*1.5*
Style/Awkward	5.7	17.1	0.122	0.182	*1.5*	0.193	*1.6*	0.180	*1.5*	0.185	*1.5*
Accuracy/Addition	0.9	40.2	0.110	0.025	*0.2*	0.017	*0.1*	0.013	*0.1*	0.018	*0.2*
Fluency/Spelling	3.6	5.1	0.107	0.071	*0.7*	0.071	*0.7*	0.059	*0.6*	0.073	*0.7*
Fluency/Punctuation	11.1	1.4	0.028	0.035	*1.2*	0.035	*1.3*	0.031	*1.1*	0.033	*1.2*
Locale/Currency format	0.4	8.8	0.011	0.010	*0.9*	0.010	*0.9*	0.010	*0.9*	0.010	*0.9*
Fluency/Inconsistency	0.8	27.5	0.011	0.036	*3.3*	0.028	*2.7*	0.026	*2.4*	0.038	*3.5*
Fluency/Register	0.4	6.5	0.008	0.008	*1.0*	0.008	*0.9*	0.008	*1.0*	0.009	*1.1*
Locale/Address format	0.3	65.7	0.008	0.025	*3.3*	0.036	*4.7*	0.033	*4.3*	0.015	*2.0*
Non-translation	0.0	100.0	0.006	0.024	*3.9*	0.021	*3.3*	0.012	*2.0*	0.029	*4.7*
Terminology/Inconsistent	0.3	16.1	0.004	0.008	*2.3*	0.007	*1.8*	0.004	*1.2*	0.010	*2.8*
Other	0.1	4.1	0.003	0.003	*0.9*	0.005	*1.7*	0.002	*0.6*	0.001	*0.4*
All accuracy	51.7	69.3	2.444	3.748	*1.5*	3.463	*1.4*	3.655	*1.5*	3.721	*1.5*
All fluency	29.8	10.5	0.535	0.593	*1.1*	0.557	*1.0*	0.517	*1.0*	0.580	*1.1*
All except acc. & fluency	18.5	41.7	0.546	0.986	*1.8*	1.005	*1.8*	0.955	*1.7*	0.891	*1.6*
All categories	100.0	46.7	3.525	5.327	*1.5*	5.025	*1.4*	5.127	*1.5*	5.192	*1.5*

Table 6: Category breakdown of MQM scores for Chinese→English for human translations (A, B), machine translations (all systems) and some of the best systems (VolcTrans, WeChat, Tencent). The ratio of system over human scores is in italics. Errors (%) report the fraction of the total error counts in a category, Major (%) report the fraction of major error for each category.

4.3 Document-error Distribution

We calculate document-level scores by averaging the segment level scores of each document.

We show the average document scores of all MT systems and all human translations (HT) for English→German in Figure 5. The translation quality of humans is very consistent over all docu-

8

Table 7 (a) English→German

Categories	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5		Rater 6	
	MQM	vs avg.	MQM	vs avg.	MQM	vs avg.	MQM	vs avg.	MQM	vs avg.	MQM	vs avg.
Accuracy	1.02	0.84	0.82	0.68	1.55	1.28	1.42	1.18	1.23	1.02	1.21	1.00
Fluency	0.26	0.96	0.34	1.27	0.32	1.18	0.28	1.04	0.19	0.70	0.23	0.86
Others	0.41	0.80	0.63	1.23	0.59	1.14	0.57	1.10	0.57	1.10	0.32	0.63
All	1.69	0.85	1.79	0.90	2.45	1.23	2.27	1.14	1.98	1.00	1.76	0.88

(b) Chinese→English

Categories	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5		Rater 6	
	MQM	vs avg.	MQM	vs avg.	MQM	vs avg.	MQM	vs avg.	MQM	vs avg.	MQM	vs avg.
Accuracy	3.34	0.96	3.26	0.94	3.31	0.95	2.51	0.72	4.57	1.31	3.91	1.12
Fluency	0.39	0.68	0.50	0.87	1.13	1.95	0.33	0.57	0.59	1.02	0.53	0.92
Others	0.70	0.78	0.75	0.83	0.85	0.94	0.66	0.74	1.11	1.24	1.32	1.47
All	4.43	0.89	4.51	0.91	5.29	1.07	3.50	0.71	6.27	1.26	5.76	1.16

Table 7: MQM per rater and category. The ratio of a rater score over the average score is in italics.

ments and gets a MQM score of around 1 which is equivalent to one minor error. This demonstrates that the translation quality of humans is consistent independent of the underlying source sentence. The distribution of MQM errors for machine translations looks much different. For some documents, MT gets very close to human performance, while for other documents the gap is clearly visible. Interestingly, all MT systems have similar problems with the same subset of documents which demonstrated that the quality of MT output is more conditioned on the actual input sentence and not only on the underlying MT system.

The MQM document-level scores for Chinese→English are shown in Figure 6. The distribution of MQM errors for the MT output looks very similar to the ones for English→German. There are documents that are more challenging for some MT systems than others. Although the document-level scores are mostly lower for human translations, the distribution looks similar to the ones from MT systems. We first suspected that the reference translations were post-edited from MT. This is not the case: these translations originate from professional translators without access to post-editing but with access to CAT tools (mem-source and translation memory). Another possible explanation is the nature of the source sentences. Most sentences come from Chinese government news pages which have a formal style that may be difficult to render in English.

4.4 Annotator Agreement and Reliability

Our annotations were performed by professional raters with MQM training. All raters were given
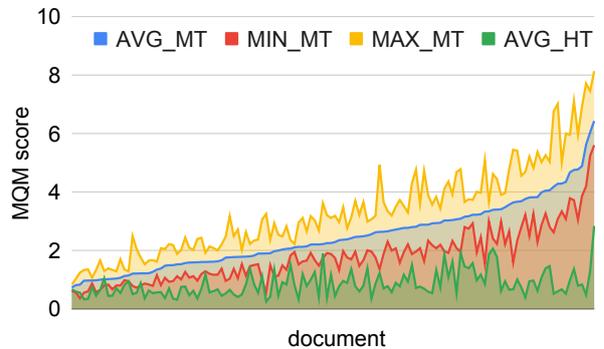

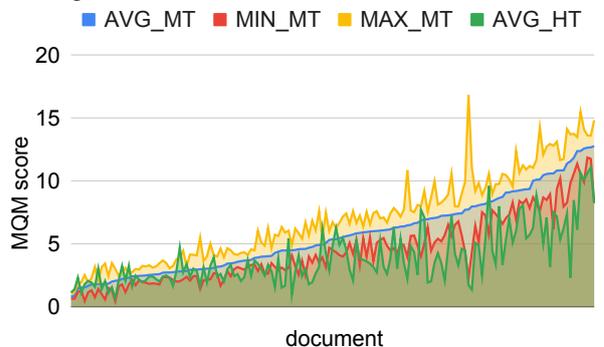
Figure 5: EnDe: Document-level MQM scores.

Figure 6: ZhEn: Document-level MQM scores.

roughly the same amount of work, with the same number of segments from each system. This setup should result in similar aggregated rater scores.

Table 7(a) reports the scores per rater aggregated over the main error categories for English→German. All raters provide scores within ±20% around the mean, with rater 3 being the most severe rater and rater 1 the most permissive. Looking at individual ratings, rater 2 rated fewer errors in accuracy categories but used the style/awkward category more for errors outside of fluency/accuracy. Conversely, rater 6 barely used this category. Differences in error rates among

raters are not severe but could be reduced with corrections from an annotation model (Paun et al., 2018), especially when working with larger annotator pools.

The rater comparison on Chinese→English in Table 7(b) reports a wider range of scores than for English→German. All raters provide scores within ±30% around the mean. This difference might be due to the greater difficulty of the translation task itself introducing more ambiguity in the labeling. In the future, it would be interesting to compare if translation between languages of different families suffer larger annotator disagreement for MQM ratings.

4.5 Number of MQM Ratings Required

Human evaluation with professional translators is more expensive than using the crowd. To keep the cost as low as possible, we compute the minimum number of ratings required to get a reliable human evaluation. We simulate new MQM rating projects by bootstrapping from the existing MQM data.[2] We compute Kendall's τ correlation of the simulated system level scores with the system level scores obtained from the full MQM data set. Note that later should be considered as the ground truth when estimating the accuracy of simulated MQM projects. See Figure 7 for the change of distributions of Kendall's τ for English→German as the number of ratings increases.
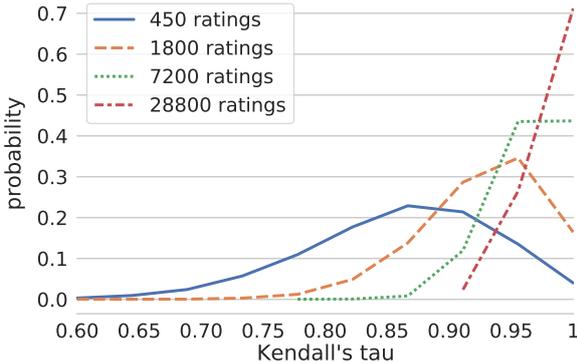

Figure 7: Distributions of Kendall's τ of system level scores for English→German. As the number of ratings increases, the distribution of Kendall's τ converges to the Dirac distribution at 1. All systems use 1 rater per sentence and 3 consecutive sentences per document. The width of 95% CI is small (< 0.02), and thus is not shown here.

Figure 8 shows the effect of different distributing schema for a fixed budget of 900 segment-level ratings. The system level scores become more accurate when limiting the number of segment-level ratings to 3 consecutive sentences in each document and thus distributing the 900 segment-level scores over more documents.
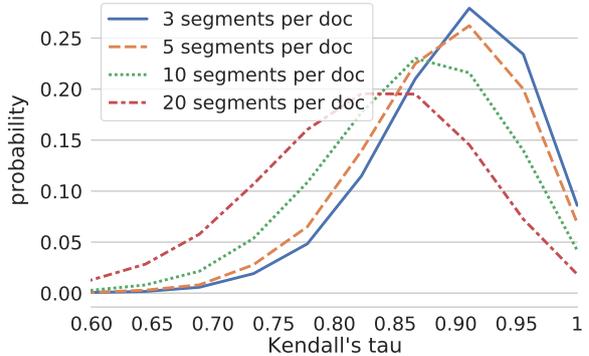

Figure 8: System-level Kendall's τ for different distribution schema of 900 segment-level ratings for English→German.

Once the items to be rated is fixed for one system, aligning the ratings across different systems makes the comparison of two system more accurate. For MQM, this means that to compare different systems, it helps to rate the same documents, and the same sentences in the corresponding documents. When possible, using the same rater(s) to rate the corresponding sentences for different systems further improves the accuracy of the comparison between systems.

Finally, we estimate the number of ratings needed for MQM on different language pairs. The estimations are for systems with 3 consecutive sentences rated per document, and 1 rating per sentence. We further align the documents and the sentences rated across systems, but we do not align raters for corresponding sentences. We estimate the minimum number of ratings required such that the expected Kendall's τ correlation with the full data set ≥ 0.9.

language pair	number of ratings required
English→German	951
Chinese→English	3720

Table 8: MQM: Number of required ratings per system to achieve Kendall's τ of 0.9
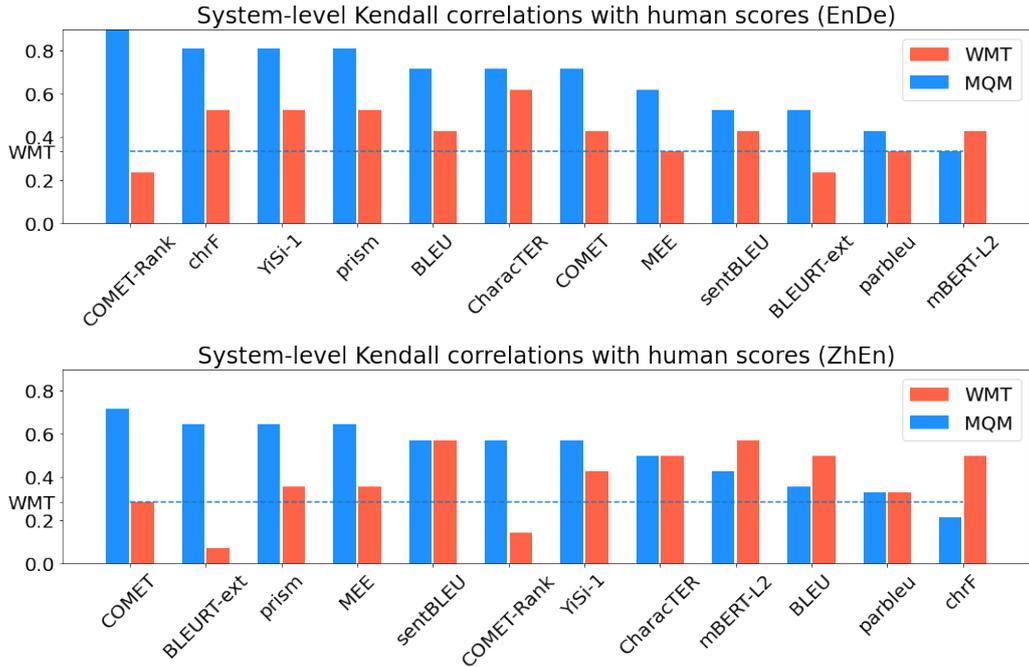
Figure 9: System-level metric performance with MQM and WMT scoring for: (a) EnDe, top panel; and (b) ZhEn, bottom panel. The horizontal blue line indicates the correlation between MQM and WMT human scores.

4.6 Impact on Automatic Evaluation

We compared the performance of automatic metrics submitted to the WMT20 Metrics Task when gold scores came from the original WMT ratings to the performance when gold scores were derived from our MQM ratings. Figure 9 shows Kendall's tau correlation for selected metrics at the system level for English→German and Chinese→English;[3] full results are in Appendix C. As would be expected from the low correlation between MQM and WMT scores, the ranking of metrics changes completely under MQM. In general, metrics that are not solely based on surface characteristics do somewhat better, though this pattern is not consistent (for example, chrF has a correlation of 0.8 for EnDe). Metrics tend to correlate better with MQM than they do with WMT, and almost all achieve better MQM correlation than WMT does (horizontal dotted line).

Table 9 shows average correlations with WMT and MQM gold scores for different granularities. At the system level, correlations are higher for MQM than WMT, and for EnDe than ZhEn. Correlations to MQM are quite good, though on aver-

Average correlations	EnDe		ZhEn	
	WMT	MQM	WMT	MQM
Pearson, sys-level	0.539	0.883	0.318	0.551
	0.23	*0.02*	*0.41*	*0.21*
Kendall, sys-level	0.436	0.637	0.309	0.443
	0.27	*0.10*	*0.42*	*0.23*
Kendall, sys-level baselines only	0.467	0.676	0.514	0.343
	0.20	*0.06*	*0.10*	*0.34*
Kendall, sys-level +human	0.387	0.123	0.426	0.159
	0.26	*0.68*	*0.20*	*0.64*
Kendall, seg-level	0.170	0.228	0.159	0.298
	0.00	*0.00*	*0.00*	*0.00*
Kendall, seg-level +human	0.159	0.161	0.157	0.276
	0.00	*0.00*	*0.00*	*0.00*

Table 9: Average correlations for metrics at different granularities. The *baselines only* result averages over BLEU, sentBLEU, TER, chrF, and chrF++; other results average over all metrics available for the given condition. The *+human* results include reference translations among outputs to be scored: one standard and one paraphrased reference for EnDe, and one standard reference for ZhEn. Numbers in italics are average p-values from two-tailed tests, indicating the probability that the observed correlation was due to chance.

age they are statistically significant only for EnDe. Interestingly, the average performance of baseline metrics is similar to the global average for all metrics in all conditions except for ZhEn WMT, where it is substantially better. Adding human transla-

---

[3]The official WMT system-level results use Pearson correlation, but since we are rating fewer systems (only 7 in the case of EnDe), Kendall is more meaningful; it also corresponds more directly to the main use case of system ranking.

tions to the outputs scored by the metrics results in a large drop in performance, especially for MQM, due to human outputs being rated unambiguously higher than MT by MQM. Segment-level correlations are generally much lower than system-level, though they are significant due to having greater support. MQM correlations are again higher than WMT at this granularity, and are higher for ZhEn than EnDe, reversing the pattern from system-level results and suggesting a potential for improved system-level metric performance through better aggregation of segment-level scores.

5 Conclusion

As part of this work, we proposed a standard MQM scoring scheme that is appropriate for high-quality MT. We used MQM to acquire ratings by professional translators for the recent WMT 2020 evaluation campaign for Chinese→English and English→German and used them as a platinum standard for comparison to different simpler evaluation methodologies and crowd worker evaluations. We release all ratings acquired in this study to encourage further research on this dataset for both human evaluation and automatic evaluation.

Our study shows that crowd-worker human evaluations (as conducted by WMT) have low correlation with MQM, and the resulting system-level rankings are quite different. This finding questions previous conclusions made on the basis of crowd-worker human evaluation, especially for high-quality MT. We further come to the surprising finding that many automatic metrics, and in particular embedding-based ones, already outperform crowd-worker human evaluation. Unlike ratings acquired by crowd-worker and ratings acquired by professional translators on simpler human evaluation methodologies, MQM labels acquired with professional translators show a large gap between the quality of human and machine generated translations. This demonstrates that MT is still far from human parity. Furthermore, we characterize the current error types in human and machine translations, highlighting which error types are responsible for the difference between the two. We hope that researchers will use this as motivation to establish more error-type specific research directions. Finally, we give recommendations of how many MQM labels are required to establish a reliable human evaluation and how these

ratings should be distributed across documents.

References

ALPAC. 1966. *Language and Machines: Computers in Translation and Linguistics; a Report*, volume 1416. National Academies.

Eleftherios Avramidis, Aljoscha Burchardt, Christian Federmann, Maja Popović, Cindy Tscherwinka, and David Vilar. 2012. Involving Language Professionals in the Evaluation of Machine Translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1127–1130, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. 2018. Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment. In *International Workshop on Spoken Language Translation*.

Ondrej Bojar, Rajen Chatterjee, Federmann Christian, Graham Yvette, Haddow Barry, Huck Matthias, Koehn Philipp, Liu Qun, Logacheva Varvara, Monz Christof, et al. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Second Conference onMachine Translation*, pages 169–214. The 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.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, and Cameron Shaw Fordyce. 2008. Proceedings of the Third Workshop on Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, and Maria Gialama. 2017. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. *AAMT*.

Lukas Fischer and Samuel Läubli. 2020. What's the Difference Between Professional Human and Machine Translation? A Blind Multi-language Study on Domain-specific MT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 215–224, Lisboa, Portugal. European Association for Machine Translation.

Marina Fomicheva et al. 2017. *The Role of Human Reference Translation in Machine Translation Evaluation*. Ph.D. thesis, Universitat Pompeu Fabra.

Mikel L Forcada, Carolina Scarton, Lucia Specia, Barry Haddow, and Alexandra Birch. 2018. Exploring Gap Filling as a Cheaper Alternative to Reading Comprehension Questionnaires when Evaluating Machine Translation for Gisting. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 192–203.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at Scale and Its Implications on MT Evaluation Biases. In *Proceedings of the Fourth Conference on Machine Translation*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be Guilty but References Are Not Innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71.

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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can Machine Translation Systems be Evaluated by the Crowd Alone? *Natural Language Engineering*, 23(1):3–30.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Translationese in Machine Translation Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint arXiv:1803.05567*.

Filip Klubička, Antonio Toral, and Víctor M Sánchez-Cartagena. 2018. Quantitative Fine-Grained Human Evaluation of Machine Translation Systems: a Case Study on English to Croatian. *Machine Translation*, 32(3):195–215.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121.

Moshe Koppel and Noam Ordan. 2011. Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 1318–1326.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral.

2020. A Set of Recommendations for Assessing Human–Machine Parity in Language Translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, pages 0455–463.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. 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.

Maja Popović. 2020. Informative Manual Evaluation of Machine Translation Output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Carolina Scarton and Lucia Specia. 2016. A Reading Comprehension Corpus for Machine Translation Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3652–3658.

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at wmt 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Belgium, Brussels. Association for Computational Linguistics.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E Banchs. 2007. Human Evaluation of Machine Translation Through Binary System Comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103.

John S White, Theresa A O'Connell, and Francis E O'Mara. 1994. The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.

Mike Zhang and Antonio Toral. 2019. The Effect of Translationese in Machine Translation Test Sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81.

A MQM Summary

The Multidimensional Quality Metrics (MQM) framework was developed in the EU QT-LaunchPad and QT21 projects (2012–2018) (www.qt21.eu). It provides a generic methodology for assessing translation quality that can be adapted to a wide range of evaluation needs. The central idea is to establish a standard hierarchy of translation *issues* (potential errors) that can be pruned or extended with new issues as required. Annotators identify issues in text at a suitable granularity, and the results are summarized using a procedure that is specific to the application.

The MQM standard (www.qt21.eu/mqm-definition) consists of a controlled vocabulary for describing issues, a scoring mechanism for aggregating annotation results, an XML formalism for describing specific *metrics* (instantiations of MQM), a set of guidelines for selecting issues, and mappings from legacy metrics to MQM. All components except the vocabulary and XML mechanism are considered suggestive, and may be modified as required. Figure 10 depicts the MQM Core issue hierarchy, intended to cover common issues arising in translated texts.

Guidelines for adapting MQM to scientific research are provided in the standard, and augmented by (MQM-usage-guidelines.pdf). The main points can be summarized as follows:

- Choose an issue hierarchy suitable to the research questions being addressed, introducing new issues as needed,[4] and pruning irrelevant issues to reduce ambiguity and cognitive load. Specify the granularity of the text units to which the issues will apply; this may range from sub-sentential spans to multi-document collections.

- If possible, use expert human translators or translators to perform annotations; three annotators per text item is recommended. Provide training in the use of the annotation tool, and guidelines for interpreting the issue hierarchy. These may be augmented with examples or decision trees, and a calibration set containing known errors can be used to assure annotator competence.

- Annotation should proceed in short segments

(30 minutes), and the allocated time should take text difficulty into account. Annotation cost is estimated to be approximately 1 USD / segment (assuming three annotators), but can be highly variable. Annotation within document context is assumed implicitly.

- Analysis can produce aggregate scores or finer-grained summaries. The specification recommends that each issue be graded with a severity: none, minor, major, or critical. Aggregate scores can weight each issue by type (the default is to weight all types equally) and by severity (recommended scores are 0, 1, 10, and 100, respectively).

B MQM for Broad-Coverage MT

Annotation

Our broad-coverage MT issue hierarchy is shown in Table 10. It is intended to be applied at the segment level by annotators with access to document context. We based it loosely on the MQM core hierarchy, with modifications established in collaboration with expert translators from our rater pool who had MQM experience. After an initial pilot run, we added several sub-categories to *Locale convention* for the sake of consistency.[5] Apart from clarifying the definitions of some categories, our main change was to add a *Non-translation* category to cover situations where identifying individual errors would be meaningless. At most one Non-translation error can be assigned to a segment, and choosing Non-translation precludes the identification of other errors in that segment.

Table 11 shows descriptions for three severity levels that raters must assign to errors independent of their category. Many MQM schemes include an additional "Critical" severity which is worse than Major, but we dropped this because its definition is often context-specific, capturing errors that are disproportionately harmful for a particular application. We felt that for broad coverage MT the distinction between Major and Critical was likely to be highly subjective, while Major errors (actual errors) would be easier to distinguish from Minor ones (imperfections). Neutral severity allows annotators to express subjective opinions about the translation without affecting its rating.

---

[4]These must not overlap semantically with issues in the controlled vocabulary.

[5]An alternative and arguably preferable strategy would have been to collapse all sub-categories for locale.
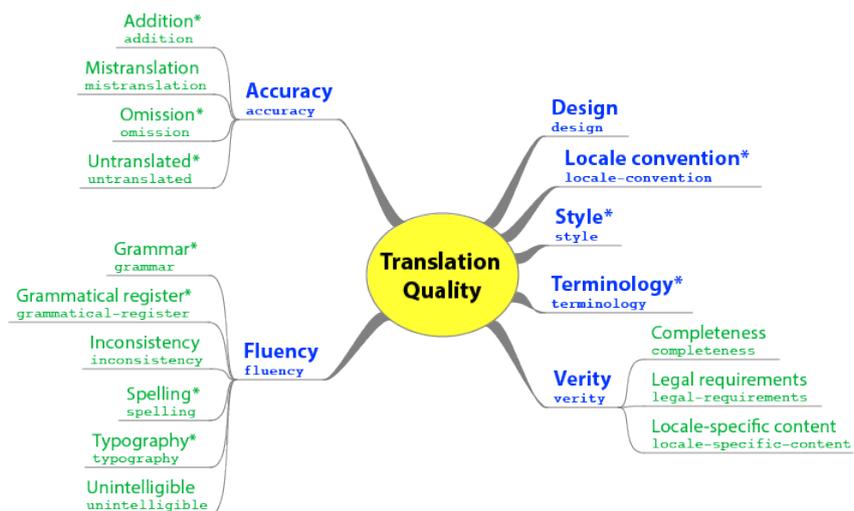
Figure 10: MQM Core issue hierarchy.

Annotator instructions are shown in Table 12. We kept these minimal because our raters were professionals with previous experience in assessing translation quality, including with MQM. There are many subtle issues that arise in error annotation, such as the correct way to translate units (eg, should 1 inch be translated as 1 Zoll, 1cm, or 2.54cm?), but we resisted the temptation to establish an extensive list of context-specific guidelines, relying instead on the judgment of our annotators. In order to temper the effect of long segments, we imposed a maximum of five errors per segment. For segments with more errors, we asked raters to identify only the five most severe. Thus we do not distinguish between segments containing five or more than five Major errors, although we do distinguish between segments with many identifiable errors and those that are categorized as entirely Non-translation. To focus our raters on careful error identification, and to provide potentially useful information for further studies, we had them highlight error spans in the text, following the conventions laid out in Table 12.

Scoring

Since we are ultimately interested in deriving scores for sentences, we require a weighting on error categories and severities. We set the weight on Minor errors to 1, and explored a range of Major error weights from 1 to 10 (the Major weight recommended in the MQM standard). For each weight combination we examined the stability of system ranking using a resampling technique. We found that a Major weight of 5 gave the best balance of stability and ability to discriminate among systems.

These weights apply to all error categories except Fluency/Punctuation and Non-translation. We assigned a weight of 0.1 for Fluency/Punctuation to reflect its mostly non-linguistic character. Decisions like the kind of quotation mark to use or the spacing between words and punctuation affect the appearance of a text but do not change its meaning. Unlike other kinds of minor errors, these are easy to correct algorithmically, so we assign them a low weight to ensure that their main role is to distinguish between systems that are equivalent in other respects. Our decision is supported by evidence from professional translators, who tend to treat minor punctuation errors as insignificant for the purpose of scoring, even when they are required to annotate them within the MQM framework. Note that this category does not include punctuation errors that render a text ungrammatical or change its meaning (eg, eliding the comma in "Let's eat, grandma"), which have the same weight as other Major errors. Source errors are ignored in our current study but give us the ability to discard badly garbled source sentences, which might be prevalent in certain genres. The singleton Non-translation category has a weight of 25, equivalent to five Major errors, the worst segment-level score possible in our annotation scheme.

Our current weighting ignores the text span of errors, as this provides little information relevant to scoring once severity and category are taken into account.

Error Category		Description
Accuracy	Addition	Translation includes information not present in the source.
	Omission	Translation is missing content from the source.
	Mistranslation	Translation does not accurately represent the source.
	Untranslated text	Source text has been left untranslated.
Fluency	Punctuation	Incorrect punctuation (for locale or style).
	Spelling	Incorrect spelling or capitalization.
	Grammar	Problems with grammar, other than orthography.
	Register	Wrong grammatical register (eg, inappropriately informal pronouns).
	Inconsistency	Internal inconsistency (not related to terminology).
	Character encoding	Characters are garbled due to incorrect encoding.
Terminology	Inappropriate for context	Terminology is non-standard or does not fit context.
	Inconsistent use	Terminology is used inconsistently.
Style	Awkward	Translation has stylistic problems.
Locale convention	Address format	Wrong format for addresses.
	Currency format	Wrong format for currency.
	Date format	Wrong format for dates.
	Name format	Wrong format for names.
	Telephone format	Wrong format for telephone numbers.
	Time format	Wrong format for time expressions.
Other		Any other issues.
Source error		An error in the source.
Non-translation		Impossible to reliably characterize the 5 most severe errors.

Table 10: MQM hierarchy.

Severity	Description
Major	Errors that may confuse or mislead the reader due to significant change in meaning or because they appear in a visible or important part of the content.
Minor	Errors that don't lead to loss of meaning and wouldn't confuse or mislead the reader but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing.
Neutral	Use to log additional information, problems or changes to be made that don't count as errors, e.g. they reflect a reviewer's choice or preferred style.

Table 11: MQM severity levels.

Table 1 summarizes our weighting scheme. The score of a segment is the sum of all errors it contains, averaged over all annotators, and ranges from 0 (perfect) to 25 (maximally bad). Segment scores are averaged to provide document- and system-level scores.

C Analysis of Metric Performance

Figure 12 shows the system-level Kendall tau correlations for all metrics from the WMT 2020 metrics task, completing the partial picture given in Figure 9. Figure 11 contains the corresponding plots for Pearson correlation. Figure 13 shows Kendall correlation for English→German for metrics using the paraphrased references available for that language pair; this substantially changes metric ranking and performance. Finally, Figure 17

shows performance when human outputs were included among the systems to be scored, resulting in lower correlations compared to MQM gold scores, and much lower correlations compared to WMT gold scores.

For segment-level correlations, we adopted the WMT "Kendall-like" measure to deal with missing and unreliable segment-level annotations in the WMT data. This discards pairwise rankings when annotations are missing or when raw scores differ by less than 25. This statistic aggregates pairwise rankings over system scores for each segment rather than working from a single global list of segment-level scores, independent of which system they pertain to. For MQM correlations, lacking a way to establish a comparable threshold, and because we expected small differences to

You will be assessing translations at the segment level, where a segment may contain one or more sentences. Each segment is aligned with a corresponding source segment, and both segments are displayed within their respective documents. Annotate segments in natural order, as if you were reading the document. You may return to revise previous segments.

Please identify all errors within each translated segment, up to a maximum of five. If there are more than five errors, identify only the five most severe. If it is not possible to reliably identify distinct errors because the translation is too badly garbled or is unrelated to the source, then mark a single *Non-translation* error that spans the entire segment.

To identify an error, highlight the relevant span of text, and select a category/sub-category and severity level from the available options. (The span of text may be in the source segment if the error is a source error or an omission.) When identifying errors, please be as fine-grained as possible. For example, if a sentence contains two words that are each mistranslated, two separate mistranslation errors should be recorded. If a single stretch of text contains multiple errors, you only need to indicate the one that is most severe. If all have the same severity, choose the first matching category listed in the error typology (eg, *Accuracy*, then *Fluency*, then *Terminology*, etc).

Please pay particular attention to document context when annotating. If a translation might be questionable on its own but is fine in the context of the document, it should not be considered erroneous; conversely, if a translation might be acceptable in some context, but not within the current document, it should be marked as wrong.

There are two special error categories: *Source error* and *Non-translation*. Source errors should be annotated separately, highlighting the relevant span in the source segment. They do not count against the 5-error limit for target errors, which should be handled in the usual way, whether or not they resulted from a source error. There can be at most one *Non-translation* error per segment, and it should span the entire segment. No other errors should be identified if *Non-Translation* is selected.

Table 12: MQM annotator guidelines

be significant, we used a threshold of 0. The results are shown in Figures 15, 16, and 17 for standard references, paraphrased references, and with human outputs included, respectively. In general, segment-level correlations are much lower than system-level, but patterns of differences between WMT and MQM correlations remain similar.
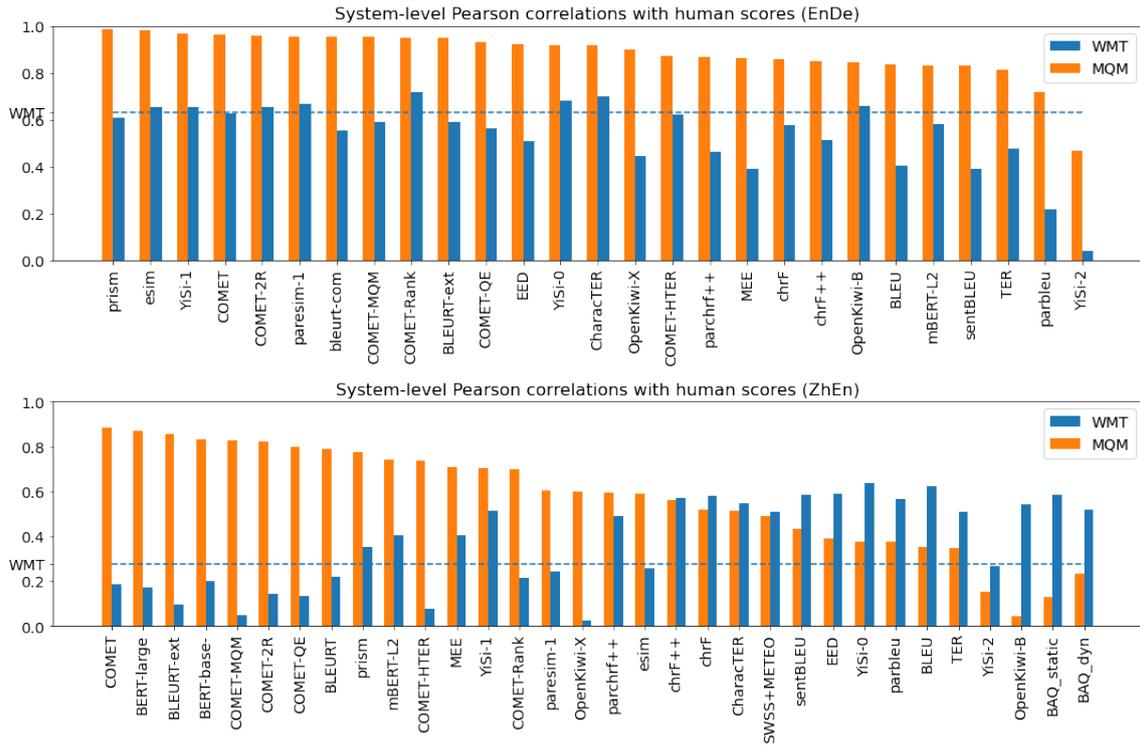
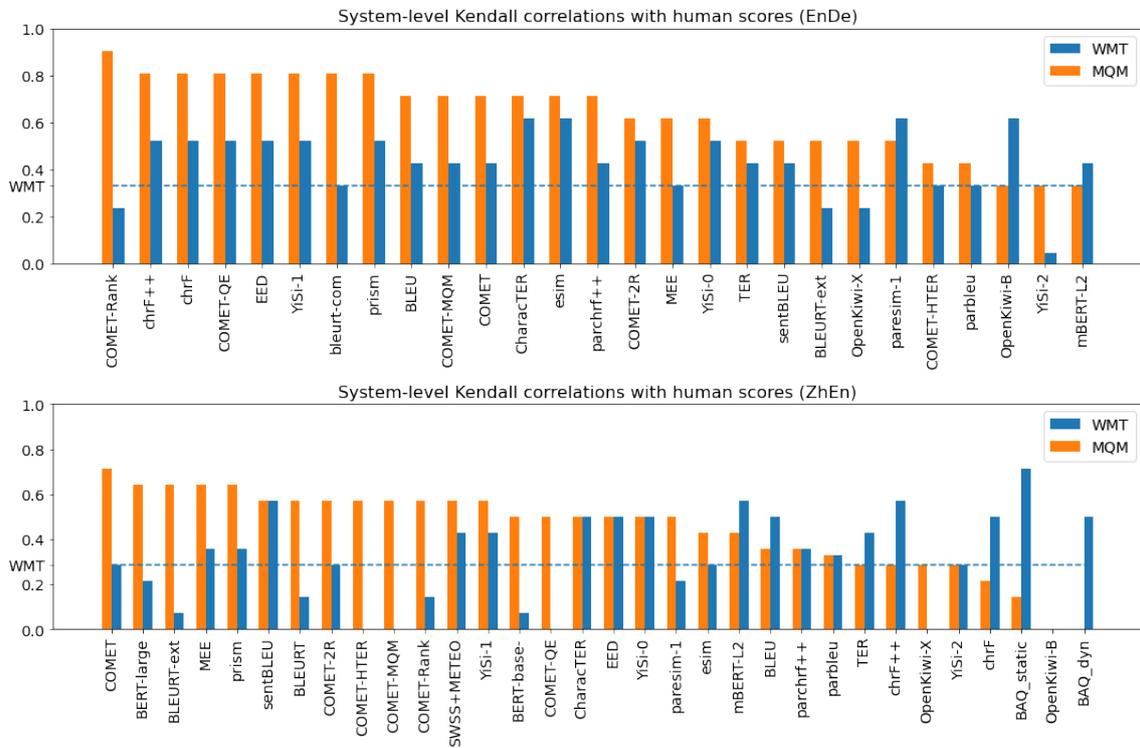Figure 11: System-level Pearson correlation with MQM and WMT scoring.

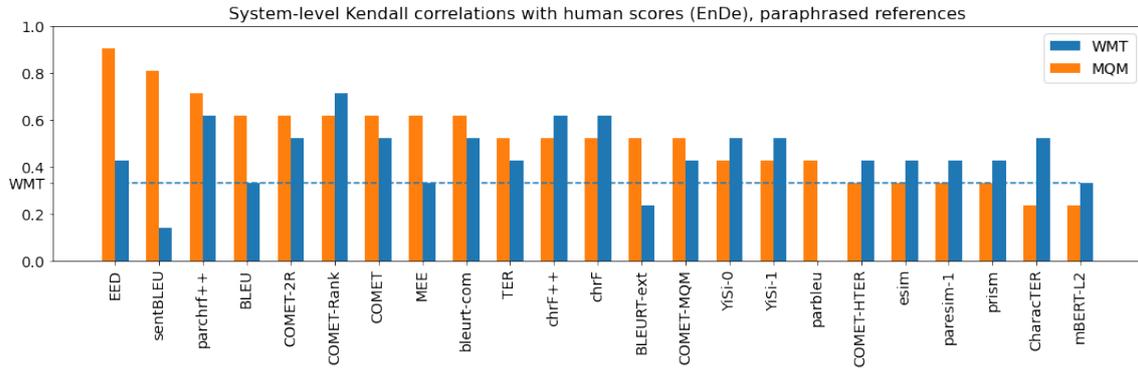Figure 12: System-level Kendall correlation with MQM and WMT scoring.

Figure 13: System-level Kendall correlation with MQM and WMT scoring when metrics use paraphrased reference.
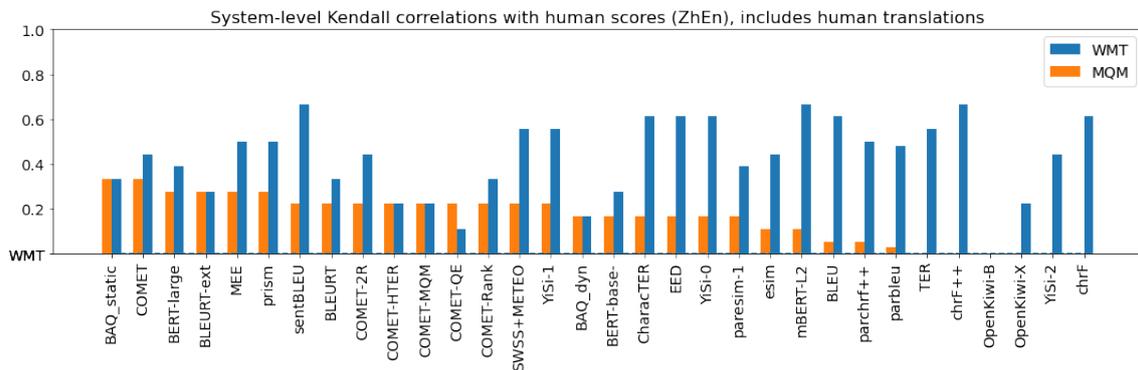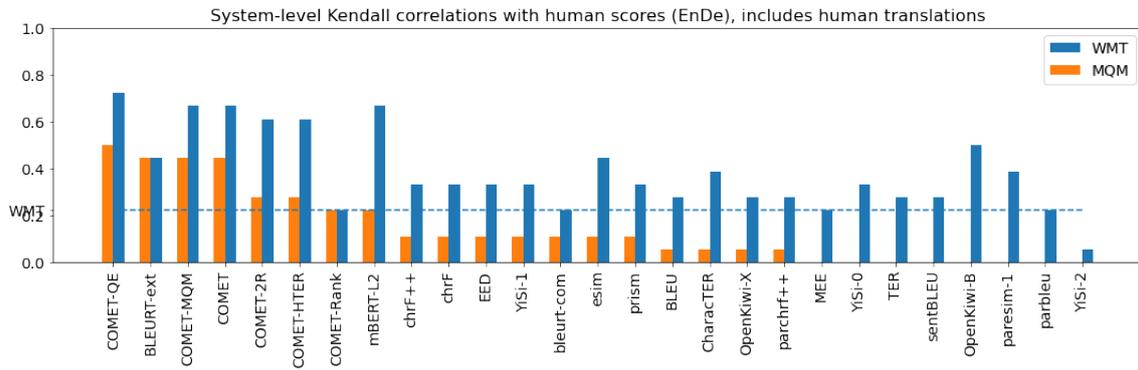

Figure 14: System-level Kendall correlation with MQM and WMT scoring when human outputs are included among systems to be scored.
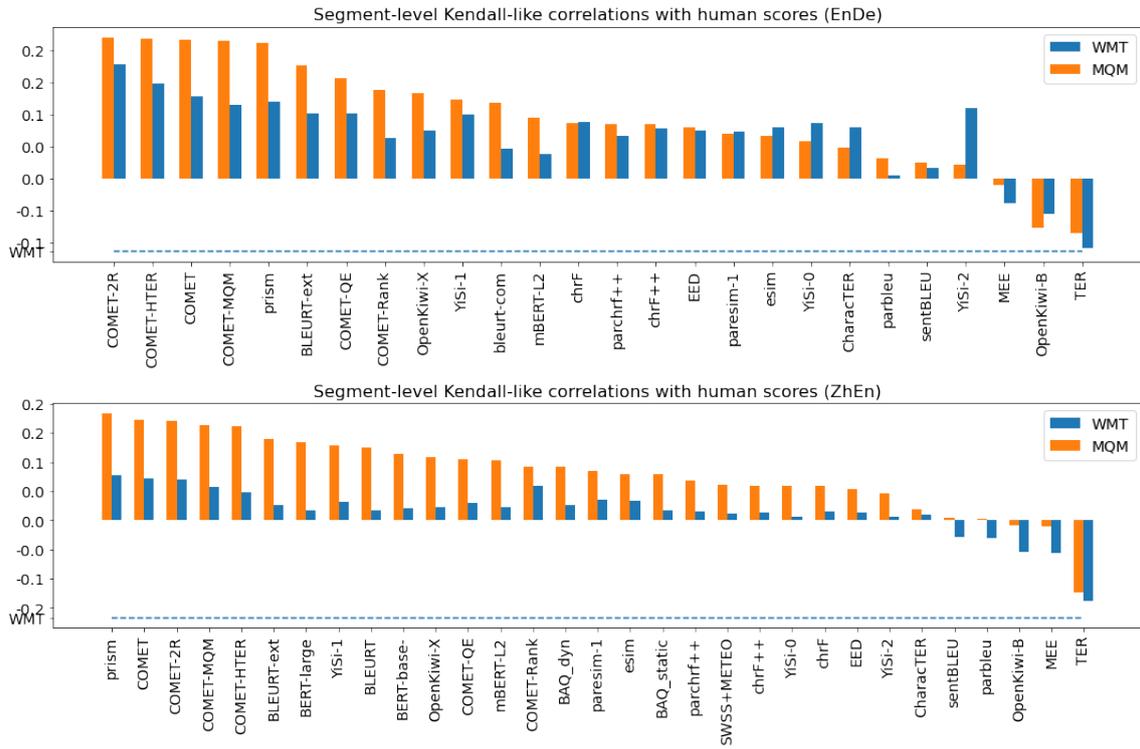
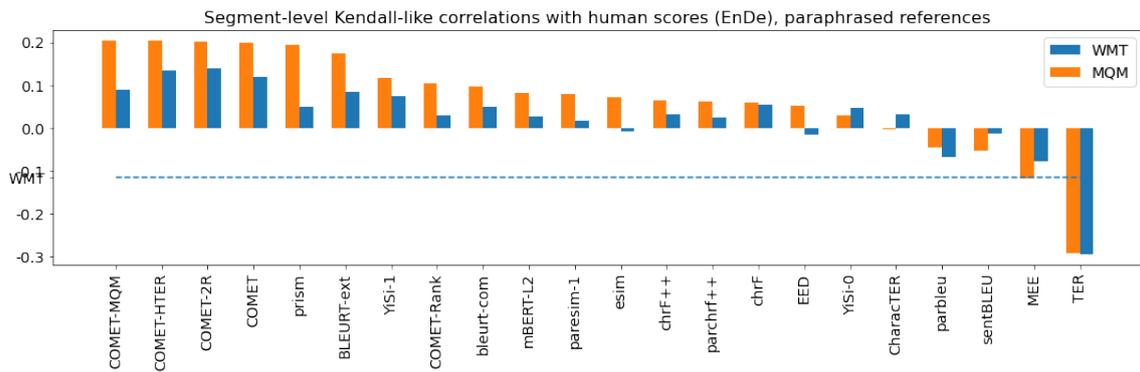Figure 15: Segment-level Kendall correlation with MQM and WMT scoring.

Figure 16: Segment-level Kendall correlation with MQM and WMT scoring when metrics use paraphrased reference.
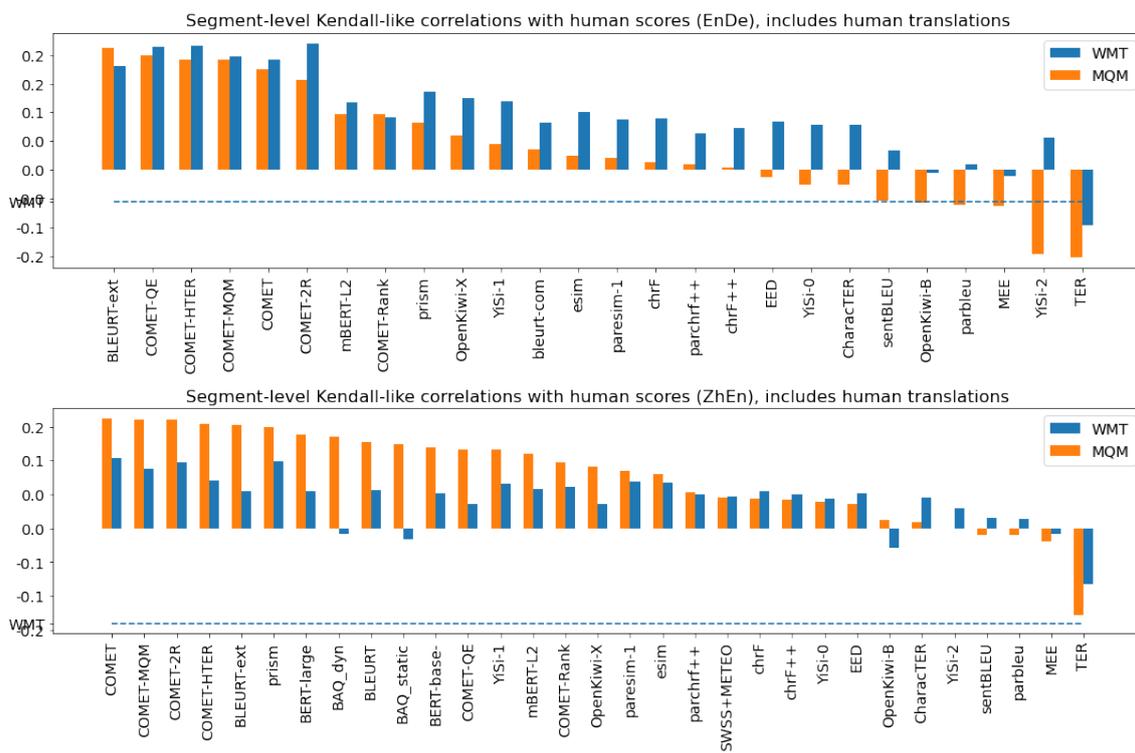
Figure 17: Segment-level Kendall correlation with MQM and WMT scoring when human outputs are included among systems to be scored.