MS-COMET: More and Better Human Judgements Improve Metric Performance

Tom Kocmi **Hitokazu Matsushita** **Christian Federmann**

Microsoft, 1 Microsoft Way, Redmond, WA 98052, USA
{tomkocmi,himatsus,chrife}@microsoft.com

Abstract

We develop two new metrics that build on top of the COMET architecture. The main contribution is collecting a ten-times larger corpus of human judgements than COMET and investigating how to filter out problematic human judgements. We propose filtering human judgements where human reference is statistically worse than machine translation. Furthermore, we average scores of all equal segments evaluated multiple times. The results comparing automatic metrics on source-based DA and MQM-style human judgement show state-of-the-art performance on a system-level pair-wise system ranking. We release both of our metrics for public use.[1]

1 Introduction

Automatic metrics for machine translation (MT) evaluation are commonly used as the primary tool for comparing the translation quality of MT systems, often without evaluating systems with the human judgement that can be expensive and time-consuming (Marie et al., 2021). Therefore, studying and developing metrics that correlate well with human judgement is critical.

There is an increasing effort in the evaluation of automatic MT metrics, leading with the annual evaluation of metrics at the WMT conference (Freitag et al., 2021b,a; Kocmi et al., 2021; Mathur et al., 2020b). Most research has focused on comparing segment-level or system-level correlations between absolute metric scores and human judgements. However, Mathur et al. (2020a) emphasize that this scenario is not identical to the everyday use of metrics, where instead, researchers and practitioners use automatic scores to compare pairs of systems. For example, when claiming a new state-of-the-art, evaluating different model architectures, and deciding whether to publish results or deploy new production systems.

In this work, we focus on training automatic metric based on COMET architecture (Rei et al., 2020) utilizing a large internal trainset of human segment-level judgements. Additionally, we evaluate the metrics in a pair-wise system-level evaluation against human judgement.

We develop two metrics: *MS-COMET* intended for reference-based evaluating systems, while *MS-COMET-QE* is designed for quality estimation or source-based evaluation. We use the suffix "-22" to differentiate the models from potential future releases.

2 Related work

There are two main categories of automatic MT metrics: (1) string-based metrics and (2) metrics using pretrained models. The former compares the coverage of various substrings between the human-generated reference and MT translations, this group includes metrics such as ChrF (Popović, 2015), BLEU (Papineni et al., 2002), or TER (Snover et al., 2006). String-based methods largely depend on the quality of reference translations. However, their advantage is that their performance is predictable as it can easily diagnose which substrings affect the score the most.

The latter category of pretrained methods consists of metrics that usually use pretrained models to evaluate the quality of MT translations given the source sentence, the human reference, or both. Evaluation metrics from this category includes COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), or BERTScore (Zhang* et al., 2020). They are not strictly dependent on the reference quality (for example, they can better evaluate synonyms or paraphrases), and many studies (Freitag et al., 2021b; Mathur et al., 2020b; Kocmi et al., 2021) showed their superiority over string-based metrics. On the other hand, their performance is influenced

[1] https://github.com/MicrosoftTranslator/MS-Comet

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 541–548
Abu Dhabi, December 7–8, 2022.

by the data on which they have been trained, which may introduce bias, and the pretrained models present a black-box problem where it is challenging to diagnose potential unexpected behavior of the metric.

A separate category of automatic metrics is whether they need a human reference for evaluation. Automatic metrics that calculate scores without the need for reference (quality estimation) open the possibility of evaluating monolingual testsets that can be tailored for a specific domain without the need to build expensive human references.

We build our metric with the architecture of COMET (Rei et al., 2020).[2] It uses the Estimator model which uses pretrained language models XLM-RoBERTa to encode source, MT hypothesis and reference in the same cross-lingual space. The model is then fine-tuned on human judgement data. We use the identical hyper-parameters as COMET.

3 Human Judgement Trainset

For training our models, we use a mix of public and internal data that we further denoise by filtering out potentially problematic human judgements.

We use the same human judgments data used to train the COMET model, i.e. WMT 2017-2019 (Barrault et al., 2019; Bojar et al., 2018, 2017). To test the quality of metrics, we use WMT 2020 (Barrault et al., 2020), WMT 2021 (Akhbardeh et al., 2021) and MQM 2021 (Freitag et al., 2021b). Furthermore, we submitted our model to WMT Metrics Shared Task 2022.

In addition to publicly available data, we use a set of internal data, described in Kocmi et al. (2021) plus newer data collected over the last year. All our internal data are collected with the use of expert annotators. We use a mix of human judgement methods: source-based Direct Assessment (srcDA) (Graham et al., 2013; Federmann, 2018), contrastive Direct Assessment (contrDA, which asks users to rate pairs of system outputs), and SQM presented at WMT General MT 2022 (which uses labeled scale). All collected labels are on a scale of 0-100, where the interface structure is the main difference for human annotators.

We use internal testsets for human judgements that have been translated with a tandem of two professional translators, following findings of Freitag et al. (2020) that high-quality reference plays an essential role in automatic evaluation.

	Langs.	Domains	Segments
All available data			6.53 M
Removed low-quality			0.79 M
Removed WMT refDA			0.35 M
Removed by averaging			2.12 M
MS-COMET	111	15	2.06 M
MS-COMET-QE	113	15	3.43 M
COMET	13	1	0.66 M

Table 1: The statistics of the training corpora and the effect of filtering in terms of unique languages on the target side, unique domains, and count of training segments used to train MS-COMET, MS-COMET-QE, and original COMET.

In contrast to publicly available data that uses only the News domain, we use a mix of fifteen domains (news, conversation, legal, medical, social, e-commerce, tech, finance, and others). The news domain is the largest domain utilizing at least half of human judgements. Our collection of human judgement data covers 113 languages in contrast to 13 on which COMET is trained. A complete list of all supported languages and counts of human judgement for the largest translation directions are in Appendix A.

Reference-less metric MS-COMET-QE is trained using all training data and removing reference translations. Additionally, many human judgments are evaluated on data that are missing human reference, which is the reason for having more training data for MS-COMET-QE.

3.1 Using raw scores instead of z-scores

The z-score has been introduced (Graham et al., 2013) to resolve an issue with different strategies annotators may apply when judging systems. For example, an overly strict annotator may harshly penalize a system from which he annotated more segments. We partly avoid this problem in our data via a better sampling technique. We sample uniformly from each evaluated system in a way that each annotator evaluates the same number of sentences from each system. Therefore, different strategies should penalize all systems similarly.

As Knowles (2021) pointed out that z-score standardization of human judgements normalizes away both inter-annotator and system quality differences, and since we do not have a mechanism to avoid normalizing away system quality differences. There-

[2]To differentiate models, we are going to use COMET to reference Rei et al. (2020) models from ours labeled as MS-COMET

fore, we decided to use raw scores (0-100) instead of the z-score standardized counterpart.

Using raw scores has the benefit that it gives final scores some meaning. For low-quality languages, we may expect scores in the lower range (0-50), while for high-quality languages, the scores generally can be higher. Z-scores only do not represent any meaning. However, we do not advocate using our metric in an absolute fashion or comparing quality across languages.

However, we want to point out that we have seen only minor improvement when training metrics using raw scores in contrast to z-scores. Therefore, this decision is mainly on a pragmatic layer.

3.2 Professional annotators only

Freitag et al. (2021a) discuss that the quality of crowd-based human judgement is suboptimal, and human evaluation should focus on expert annotators. Professional annotators collect our internal human labels. However, data from WMT are collected in two different setups when one uses crowd-workers.

The language pairs that are from English or not containing English are collected with semi-professional to professional annotators and using source-based DA, which avoids reference bias. On the other hand, all into English language pairs are collected with crowd-workers with reference-based DA. For this reason, we decided to remove all WMT reference-based DA human judgement from our datasets, and therefore, we use only internal into-English human assessments.

3.3 Averaging same human judgement

In our data, many human judgment campaigns evaluate identical triplets *(source, hypothesis, reference)* in different campaigns. This happens when we compare identical baseline system across different campaigns or when a candidate system from the earlier campaign is later evaluated as a baseline system.

We notice that human scores fluctuate every time each triplet is evaluated. We have decided to average scores for all identical triplets to normalize the noise and balance the trainset. Averaging equal scores improved the performance of the metric.

We also experimented with taking a median of the scores, but the results have been a bit worse than averaging.

3.4 Removing low-quality human judgements

In our human annotation campaigns, we often include human reference translation as another system to measure how close MT systems are to human reference. However, scoring human references can also be used as a sanity check for the quality of campaigns or human references. Whenever we see a campaign where human reference is worse than the MT system, it suggests one of the following three scenarios: human reference contains error translations, human judgement is too noisy or misleading, or the MT system performs better than human translators. If we assume that MT systems are not outperforming human translators, a lower human reference score suggests either broken reference translation or a noisy campaign. Neither of these two outcomes is desirable for fine-tuning automatic metrics.

Therefore, we remove all campaigns containing human reference as an additional system, where any of the systems is statistically significantly better than human translation under the Mann–Whitney U test and alpha threshold of 5%.

4 Evaluation

Evaluation of automatic metrics is a challenging task investigated in a yearly WMT Metrics shared task (Freitag et al., 2021b). However, there is no community-agreed testset or evaluation method for comparing with humans that are considered gold standards.

There are different dimensions how to evaluate automatic MT metrics. Let's summarize the main differing points:

- **Human annotation methods** - source-based direct assessment (DA) (Graham et al., 2013), reference-based DA (Graham et al., 2013), contrastive DA (Akhbardeh et al., 2021), Multidimensional Quality Metrics (MQM) (Freitag et al., 2021a)

- **Granularity of evaluation** - evaluating correlation with human on a segment-level or system-level

- **Correlation method** - correlation of absolute values (Pearson or Kendall-like, Mathur et al., 2020b) or correlations in pairwise approach (pairwise accuracy, Kocmi et al., 2021; Mathur et al., 2020a

- **Usage of unlabeled part of testset** - human judgment often evaluates only a subset of the testset. Metrics can use the remaining unlabelled segments (especially for system-level setup)
- **Normalize human behavior** - use raw human scores or normalize them with z-score standardization
- **Evaluating human reference** - if additional human translated references should be evaluated like one of the systems (Freitag et al., 2021b)
- **Evaluating outlier systems** - absolute value correlations via Pearson are sensitive to outliers, therefore Mathur et al. (2020a) recommends removing outlier systems from evaluation.

The list is incomplete as there are other nuances, such as removing outlier systems, using only statistically significant pairs of systems, underlying quality of human judgement, etc.

Evaluating all combinations of approaches is not reasonable. Therefore we mainly follow the approach defined by Kocmi et al. (2021) and also used by WMT Metrics 2021 (Freitag et al., 2021b).

Here is a list of constraints for the evaluation:

- We use only testsets produced by professional annotators as described in Section 3.2. Thus, we do not evaluate over reference-based DA.
- We focus on a system-level pairwise setup as the important use-case for automatic metrics (Kocmi et al., 2021). Thus we do not evaluate absolute value correlations with humans. Furthermore, this avoids the problem with outlier systems.
- We use only segments that have been evaluated by humans (unlabelled segments of testsets are not used).
- We use z-score normalization mainly to be comparable with past work. However, we do not consider z-score as a good standardization approach.
- We do not remove additional human references from the evaluation as metrics should be able to evaluate any translation (not only those produced with current MT systems).

	LPs	System pairs	Method
WMT20	8	565	srcDA
WMT21	9	1000	srcDA
WMT21-contr	3	198	contrDA
MQM21-news	3	301	MQM
MQM21-ted	3	247	MQM

Table 2: The statistics of human judgement sets are used for testing automatic metrics.

4.1 Evaluation methodology

We use system-level pairwise accuracy as introduced by Kocmi et al. (2021), which evaluates how often metric agrees on the ranking of two systems with human rank:

$$\text{Accuracy} = \frac{|\text{sign}(\text{metric}\Delta) = \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|}$$

We use implementation by Freitag et al. (2021b); therefore, results on the MQM21 testset agree with their findings. We use bootstrap resampling to calculate which metrics are not significantly outperformed by the winning metric with an alpha threshold of 0.05.

To test automatic metrics, we use publicly available data from different sources. You can find statistics in Table 2.

- **WMT20** and **WMT21** - we use source-based DA from Barrault et al. (2020) and Akhbardeh et al. (2021)
- **WMT21-contr** - we use contrastive DA from Akhbardeh et al. (2021). This is the only source of truly pairwise human judgements, where annotators see the outputs of two systems next to each other. We collect those pairs of systems evaluated to each other.
- **MQM21-news** and **MQM21-ted**- we MQM data from Freitag et al. (2021b), both testsets evaluate same set of systems but over different domains.

Additionally, we combine **all** testsets to calculate pairwise accuracy across all system pairs, simply by counting all system pairs where the metric agrees with human overall evaluated system pairs in all testsets.

4.2 Evaluated automatic metrics

We train two metrics MS-COMET trained with human-produced references and MS-COMET-QE

n	All 2311 ↓	WMT20 565	WMT21 1000	WMT21-contr 198	MQM21-news 301	MQM21-ted 247
MS-COMET-22	**0.826 (1)**	**0.892 (1)**	**0.864 (1)**	0.722 (2)	0.714 (4)	**0.745 (3)**
MS-COMET-QE-22	**0.821 (2)**	0.873 (2)	0.847 (2)	**0.808 (1)**	0.734 (2)	0.713 (6)
Bleurt	**0.820 (3)**	0.869 (3)	**0.864 (1)**	0.702 (3)	0.718 (3)	**0.749 (2)**
COMET	0.816 (4)	0.869 (3)	**0.864 (1)**	0.677 (5)	0.678 (5)	**0.781 (1)**
COMET-QE	0.800 (5)	0.848 (6)	0.839 (3)	0.692 (4)	**0.774 (1)**	0.652 (7)
BERTScore	0.790 (6)	0.853 (5)	0.836 (4)	0.722 (2)	0.621 (6)	0.721 (5)
chrF	0.770 (7)	0.857 (4)	0.793 (5)	0.702 (3)	0.621 (6)	0.713 (6)
BLEU	0.688 (8)	0.848 (6)	0.622 (7)	0.601 (7)	0.618 (7)	0.741 (4)
TER	0.669 (9)	0.766 (7)	0.657 (6)	0.616 (6)	0.585 (8)	0.636 (8)

Table 3: The main results for pairwise accuracy in a system-level setting. The bold scores represent metrics that are not statistically different from the winning metric with a 0.05 alpha level. The numbers in brackets show the rank of metrics. The "n" represents the number of system pairs in each evaluation.

trained only with sources and MT hypothesis. We use identical hyper-parameters as the original COMET model (Rei et al., 2020), and the models are trained for precisely four epochs.

We compare our metrics to publicly available metrics, and either have the highest correlation with humans - COMET, BLEURT, and BERTScore (Kocmi et al., 2021; Freitag et al., 2021a) or are widely used in MT field (BLEU, ChrF, TER). We use default parameters and models for each of them, specifically:

For BLEU (Papineni et al., 2002), ChrF (Popović, 2015), and TER (Snover et al., 2006), we use SacreBLEU implementation `https://github.com/mjpost/sacrebleu/` (Post, 2018) version 2.0.1. We use the "mteval-v13a" tokenizer for all language pairs except for Chinese and Japanese, which use their separate tokenizer, as is recommended.

For BERTScore (Zhang* et al., 2020), we use `https://github.com/Tiiiger/bert_score` version 0.3.11.

For BLEURT (Sellam et al., 2020), we use the recommended model "bleurt-20" and implementation `https://github.com/google-research/bleurt`.

For COMET (Rei et al., 2020), we use recommended model "wmt20-comet-da" and for COMET-QE we use "wmt21-comet-qe-mqm". The implementation is `https://github.com/Unbabel/COMET` in version 1.1.0.

5 Results

The results for the pairwise system-level scenario are in Table 3. The results over 2311 system pairs show that both our metrics outperform all other state-of-the-art metrics, with only Bleurt not being statistically worse than our metrics.

n	23595
MS-COMET-QE-22	**0.597 (1)**
COMET-QE	**0.596 (2)**
MS-COMET-22	**0.594 (3)**
Bleurt	**0.593 (4)**
COMET	0.586 (5)
BERTScore	0.567 (6)
chrF	0.557 (7)
TER	0.536 (8)
sentBLEU	0.535 (9)

Table 4: The results for pairwise accuracy in a segment-level setting over *WMT21-contr* testset. The "n" represent a number of segment pairs used in the evaluation.

The results over individual testsets show that our metrics are ranked among the top-performing metrics. Interestingly, *MQM21-news* domain seems to be easier for Quality Estimation metrics, while *MQM21-ted* shows the opposite direction. These results are interesting as the underlying systems are identical except for additional human reference.

Lastly, our metrics win in the *WMT21-contr* testset. This is the only genuinely pairwise testset where annotators saw systems next to each other while evaluating them.

Although we focus on a system-level evaluation, we evaluate how metrics perform in a segment-level setting for completeness. We use the testset *WMT21-contr* to calculate accuracies in the same fashion as for system-level scenario, but taking

pairs of segment annotations instead of system-level scores. The segment-level results in Table 4 show that our metrics, COMET-QE, and Bleurt are in the winning cluster outperforming other metrics.

6 Conclusion

We have investigated the training COMET model with a larger corpus of human judgements covering multiple domains and 113 languages.

We employed several steps of filtering low-quality or repetitive human judgement.

With those data, we trained two metrics: MS-COMET-22 and MS-COMET-QE-22, that outperform other current MT metrics on a pair-wise system-level decision task.

We release the metrics for public use.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

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.

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, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, 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.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

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.

Rebecca Knowles. 2021. On the stability of system rankings at WMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 464–477, Online. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A List of languages

Our collection of human judgements covers 113 languages, language variants, or writing systems. Here is the complete list. Note that XLM-Roberta does not support some languages:

Afrikaans, Albanian, Amharic, Arabic, Armenian, Assamese, Azeri, Bangla, Bashkir, Basque, Bosnian, Bulgarian, Burmese, Catalan, Central Kurdish, Chinese (Literary), Chinese (People's Republic of China), Chinese (Taiwan), Chinese Yue, Chuvash, Classic Chinese (Simplified), Croatian, Czech, Danish, Dari, Divehi, Dutch, English, Estonian, Faroese, Fijian, Filipino, Finnish, French, French (Canada), Galician, Georgian, German, Greek, Gujarati, Haitian Creole, Hebrew, Hindi, Hmong, Hungarian, Icelandic, Indonesian, Inuktitut, Inuktitut (Latin), Inuinnaqtun, Irish, isiZulu, Italian, Japanese, Kannada, Kazakh, Khmer, Kiswahili, Korean, Kurdish, Kyrgyz, Lao, Latvian, Lithuanian, Macedonian, Malagasy, Malay, Malay Standard, Malayalam, Maltese, Maori, Marathi, Mongolian, Mongolian (Cyrillic), Nepali, Norwegian, Odia, Otomi, Pashto, Persian, Polish, Portuguese (Brazil), Portuguese (Portugal), Punjabi, Romanian, Russian, Samoan, Serbian (Cyrillic), Serbian (Latin), Slovak, Slovenian, Somali, Spanish, Swedish, Tahitian, Tajik, Tajiki, Tamil, Tatar, Telugu, Thai, Tibetan, Tigrinya, Tongan, Turkish, Turkmen, Ukrainian, Upper Sorbian, Urdu, Uyghur, Uzbek, Vietnamese, Welsh.

Furthermore, our human judgement data are not balanced. In some translation directions, we have more human-labeled data than in others. Table 5 shows the largest forty translation directions in our training data corpus.

	Mono	With ref
English - German	175k	103k
English - Chinese	117k	80k
English - Czech	93k	71k
English - Russian	92k	72k
English - French	66k	36k
Chinese - English	63k	33k
German - English	60k	28k
Japanese - English	57k	35k
English - Japanese	55k	30k
English - Spanish	54k	27k
English - Dutch	52k	27k
French - English	50k	32k
English - Italian	50k	24k
Spanish - English	48k	29k
English - Finnish	45k	38k
Italian - English	44k	25k
English - Polish	43k	23k
Korean - English	39k	25k
English - Portuguese	38k	22k
English - Turkish	37k	24k
English - Korean	36k	20k
Polish - English	35k	19k
Czech - English	35k	18k
English - Hindi	34k	18k
English - Arabic	34k	17k
Dutch - English	33k	19k
Arabic - English	32k	16k
Russian - English	32k	17k
English - Estonian	28k	21k
English - Lithuanian	27k	18k
Hindi - English	25k	15k
Greek - English	25k	15k
English - Swedish	24k	13k
Turkish - English	23k	14k
English - Danish	21k	11k
Portuguese - English	21k	12k
English - Romanian	21k	14k
Swedish - English	21k	8k
Romanian - English	21k	16k
English - Slovak	21k	12k

Table 5: The number of human judgement for the forty largest translation directions. The counts represent data on the final filtered training set, where "Mono" are dataset counts for MS-COMET-QE and "With ref" are for MS-COMET.