# Data Sampling and (In)stability in Machine Translation Evaluation

**Chi-kiu Lo** and **Rebecca Knowles**
NRC-CNRC
Digital Technologies Research Centre
National Research Council Canada
{chikiu.lo|rebecca.knowles}@nrc-cnrc.gc.ca

## Abstract

We analyze the different data sampling approaches used in selecting data for human evaluation and ranking of machine translation systems at the highly influential Conference on Machine Translation (WMT). By using automatic evaluation metrics, we are able to focus on the impact of the data sampling procedure as separate from questions about human annotator consistency. We provide evidence that the latest data sampling approach used at WMT skews the annotated data toward shorter documents, not necessarily representative of the full test set. Lastly, we examine a new data sampling method that uses the available labour budget to sample data in a more representative manner, with the goals of improving representation of various document lengths in the sample and producing more stable rankings of system translation quality.

## 1 Introduction

Human evaluation of machine translation (MT) quality is very labour-intensive, meaning that it is not always possible to have full test sets annotated by human evaluators, for example in large-scale shared tasks like the News/General MT Task at the Conference on Machine Translation (WMT) (Kocmi et al., 2022). The typical practice in such situations is sampling a subset of the test set for human annotation to estimate the quality of MT systems; the rankings of MT systems computed from this are expected to be stable and representative of the full test set. However, in practice, inconsistency and instability in system rankings are observed in human evaluation and are often blamed on human annotator inconsistency. Thus, much of the focus on MT human evaluation is on denoising and calibrating human annotations, but there are other sources of error in the data sampling process orthogonal to annotator behavior.

Throughout WMT's history, the design of data sampling methods has been tangled up with experiments on human assessment collection approaches. In the early years when constrastive adequacy/fluency judgments and relative ranking were used (Koehn and Monz, 2006; Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015), the nature of the assessment method (i.e., comparing system outputs directly) ensured that there were overlaps in the segments annotated for the various systems. However, when direct assessment (Graham et al., 2013, 2014, 2016) was introduced (Bojar et al., 2016), the annotation subsets were selected independently per system, which is expected to produce consistent rankings (assuming sufficient annotations). In practice, some years have seen language pairs with very low annotation (sampling coverage of 12.5%) opening the door to scenarios where MT systems could be evaluated and ranked on disjoint sets of sentences, raising questions about fairness and consistency.

As WMT added document context to human evaluation (Barrault et al., 2019), segments for annotation were sampled at the document level (rather than at segment level). This introduced another source of instability and error by limiting the representation of topics and vocabulary in annotated samples, introducing systematic bias in the samples (e.g., document length), and in some cases even preventing some documents from being sampled (Knowles, 2021; Knowles and Lo, 2022). Although Miller et al. (2020) show that system performance can be resilient to adaptive overfitting against the frequently reused evaluation sets in other NLP task, they also show that NLP models' robustness to distribution shift remains a challenge. As MT research moves toward document level, document length may be related to the difficulty of the translation task. The data sampled for gold standard human evaluation should reflect the full test set to support fair system rankings and further analysis of distribution shift effects.

As discussed, sample size or coverage is a major



*Findings of the Association for Computational Linguistics: ACL 2023*, pages 13064–13074

factor in how representative the annotation subset is; with high coverage, other sources of error are less likely to cause problems, with low coverage, the errors may have compounding effects on ranking stability. However, sample size is an aspect that may be tightly constrained (e.g., due to funding).

In this paper, we study the data sampling methods and resulting instability of system rankings in WMT News/General MT task over the past four years. To focus on data sampling – separate from questions about human annotator consistency – we use automatic MT evaluation metrics to generate these system rankings. We show that system ranking consistency and representation of documents lengths in the sample can be improved by a new data sampling method that uses the available labour budget and balances the desires for document context and representativeness of the sample.

## 2 Data Sampling Methods

We divide our discussion of data sampling methods into two orthogonal components: 1) whether the subset of data annotated for each systems is sampled independently per system or once per test set and 2) how the sampling is performed.

### 2.1 Matching Subsets

One option for annotation is to sample a subset of data for annotation once from the test set, and then annotate each system's output over this fixed subset; we call this the *matching* subset condition. The alternative, used at WMT until recently, is to randomly sample data from each system's output independently. In extreme cases this can mean that there are no segments that have been annotated for all systems (Knowles and Lo, 2022). Mismatching annotation subsets appears to be more problematic when full documents are sampled than when data is sampled at the segment level (Knowles, 2021).

We argue in favour of using matching subsets, both because of the risk with mismatched subsets of introducing error into the rankings (i.e., by scoring one system on an easier subset of data) and because it offers opportunities to use statistical tests that rely on paired samples. In this paper, the simulation experiment of our proposed sampling approach is done with matching subsets of data.

### 2.2 Sampling Approach

Orthogonal to this question of matching or mismatched subsets for annotation is the question of how to sample the data. We describe three approaches that have been used at WMT and the approach we examine in this work, considering advantages and disadvantages. All suffer when there is low coverage. We briefly discuss some topics of user interface, but mostly leave that aside.

#### 2.2.1 Segment-level (SL)

Sampling at the segment level (i.e., randomly selecting segments without regard to document boundaries) has the advantage of better test set representation, especially with high coverage. The main disadvantage is the lack of document context, which is considered important for distinguishing high-quality machine translation from human translation (Läubli et al., 2018; Toral et al., 2018).[1]

#### 2.2.2 Whole document (WD)

This approach involves sampling whole documents; particularly at low coverage this may not be representative. On the other hand, it has the advantage of full document context, so there are no additional requirements of the annotation interface to incorporate context beyond the sample. In the sampling used at WMT, there has been a limit on document size; on occasion this has meant that large portions of the test set could not be sampled.

#### 2.2.3 Document Fixed Snippet (DF)

The WMT 2022 General Task (Kocmi et al., 2022) attempted a middle ground between segment-level and whole document sampling, sampling snippets of up to 10 contiguous segments (shorter snippets were only drawn when the whole document was shorter than 10 sentences).[2] The aim of this was to cover a broader range of documents while still maintaining document context. Additionally, in the user interface, annotators were shown preceding context of up to 10 snippets.

#### 2.2.4 Document Budgeted Snippet (DB)

In this proposed approach, a fixed budget is set in advance, and then snippets are sampled from documents proportional to the budget. That is, if there is budget to annotate 40% of the data (not including quality assurance or interannotator agreement annotations, which we do not discuss in this work),

[1] While this could be presented in the annotation interface, it would likely increase annotation time due to expanding the amount of data required to be read by the annotator.

[2] See Kocmi et al. (2022) and the linked repository (`https://github.com/wmt-conference/wmt22-news-systems`). Note that in some cases, longer documents may have been sampled first.

snippets corresponding to 40% of document length will be sampled. Where the document is too small or does not divide evenly, the document will be sampled (or the length of the snippet rounded up or down) to produce the correct number of snippets and snippet lengths in expectation.[3] This approach operates under the assumption that we wish to cover a wide range of documents, perform annotations with context, and produce a representative sample. In effect, this should produce document coverage percentages roughly equivalent to segment-level sampling, but with contiguous rather than discontiguous segments. For test sets with extremely long documents, this could be problematic for some annotation user interfaces.

## 3 Simulation

We show the effects of different sampling strategies by scoring segments in the sampled subset and the full test set with automatic metrics and comparing system rankings between the two.

### 3.1 Data and setup

We use data collected at the WMT News/General shared tasks from 2019 to 2022 (Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022) and organized in the MT Metrics Eval package.[4] The MT Metrics Eval package includes all scores from baseline and participating MT evaluation metrics in the Metrics shared task (Ma et al., 2019; Mathur et al., 2020; Freitag et al., 2021, 2022), covering all segments of all MT systems in WMT News/General shared tasks. It also contains complete information about which segments of each MT system were annotated, allowing us to approximate the coverage budget (without access to the actual sampling code, which was not available at the time of submission).[5] Each sampling method (using both the exact data annotated at WMT and simulations) is compared against our proposed document budgeted snippet (DB) sampling method in sumulations. All simulations are run 13 times with 13 as the random seed and we are reporting the worst, best, and median stability.

Automatic metrics are used for simulation because we can obtain and compare the system rankings between the full test set and multiple runs of different sampling approaches easily at minimal cost, testing the effects of sampling separately from annotator consistency. We compute rankings by averaging the segment-level metric scores over the sampled subset for each system. We focus on the four automatic metrics that participated in all or most of WMT19-22 Metrics shared tasks: chrF (Popović, 2015), COMET-20 (Rei et al., 2020), sentBLEU (Chen and Cherry, 2014) and YiSi-1 (Lo, 2019). sentBLEU is the sentence-level BLEU (Papineni et al., 2002), which is based on the precision of n-grams between the MT output and the reference weighted by a brevity penalty. chrF uses character n-grams, instead of word n-grams, and considers both precision and recall between the MT output and the reference. YiSi-1 measures the semantic similarity between a machine translation and human references by aggregating the IDF-weighted lexical semantic similarities based on the contextual embeddings extracted from pre-trained language models. COMET-20 is the 2020 version of COMET, which is a learnt metric fine-tuned to produce a z-standardized DA for a given translation by comparing its representation to source and reference embeddings. Though the correlations of these four metrics with human judgment on translation quality vary, it does not affect the simulation validity because we compare subset/full test set rankings from the same metric.

### 3.2 Evaluation metric

If a sampled subset were perfectly representative of the full test set, the ranking of systems computed by averaging over the segment-level scores in the subset would be identical to that obtained by averaging over the full test set. Exact scores might change, but the relative ranking of systems would be the same. We follow Knowles (2021) to use the number of language pairs where the system ranking changed to analyze the instability of human evaluation. We choose ranking change rather than cluster change because of our use of Metrics data, which ignores clusterings and focuses only on ranking. Note that we are not able to use Spear-

[3] For example, since it is not possible to sample 40% of the sentences of a document containing only one sentence, such a document would be sampled only with 40% probability (in expectation, sampling 40% of documents of length 1).

[4] https://github.com/google-research/mt-metrics-eval Using the version at commit: bdda529ce4fae9cec8156ea8a0abd94fe1b85988

[5] This may be a slight underestimate of the budget, as it does not account for duplicate annotations of the same segment, and in the segment sampling it may be a slight overestimate because identical output across different systems could be annotated just once. Appendix A summarizes average test set coverage of the sampled subsets for each translation direction in WMT 2019-2022.

| | COMET-20 | YiSi-1 | sentBLEU | chrF |
|---|---|---|---|---|
| Segment-level (Non-Matching Subsets) | | | | |
| Sampled at WMT | | | | |
| WMT19 | – | 5/6 | 5/6 | 5/6 |
| WMT20 | 2/3 | 1/3 | 1/3 | 1/3 |
| **total** | **2/3** | **6/9** | **6/9** | **6/9** |
| Simulation | | | | |
| -best | 1/3 | 3/9 | 3/9 | 3/9 |
| -median | 1/3 | 5/9 | 6/9 | 6/9 |
| -worst | 2/3 | 7/9 | 7/9 | 8/9 |
| Document Budgeted Snippet (Matching Subsets) | | | | |
| -best | 1/3 | 5/9 | 4/9 | 4/9 |
| -median | 1/3 | 6/9 | 6/9 | 7/9 |
| -worst | 2/3 | 7/9 | 7/9 | 9/9 |

Table 1: Effect of data sampling methods comparing on the data from translation directions in WMT19 and WMT20 that used segment-level sampling approach and runs of document budgeted snippet approach. Values indicate the fraction of translation directions that had changes in rank. The top section shows the real WMT results. For the simulation of WMT (middle section) and the document budgeted snippet approach (bottom section), multiple runs of simulation have been done and those with the best (min.), median and worst (max.) number changes are reported.

| | COMET-20 | YiSi-1 | sentBLEU | chrF |
|---|---|---|---|---|
| Whole Document (Non-Matching Subsets) | | | | |
| Sampled at WMT | | | | |
| WMT19 | – | 7/8 | 8/8 | 6/8 |
| WMT20 | 12/15 | 13/15 | 13/15 | 13/15 |
| WMT21 | 9/13 | 9/13 | 10/13 | 9/13 |
| WMT22 | 5/7 | 5/7 | 6/7 | 6/7 |
| **total** | **26/35** | **34/43** | **37/43** | **34/43** |
| Simulation | | | | |
| -best | 25/35 | 28/43 | 32/43 | 31/43 |
| -median | 27/35 | 34/43 | 35/43 | 33/43 |
| -worst | 29/35 | 36/43 | 37/43 | 37/43 |
| Document Budgeted Snippet (Matching Subsets) | | | | |
| -best | 13/35 | 15/43 | 25/43 | 16/43 |
| -median | 18/35 | 18/43 | 27/43 | 20/43 |
| -worst | 21/35 | 21/43 | 32/43 | 23/43 |

Table 2: Effect of data sampling methods comparing translation directions that used whole document sampling and runs of document budgeted snippet sampling.

| Samp. | COMET-20 | YiSi-1 | sentBLEU | chrF |
|---|---|---|---|---|
| Document Fixed Snippet (Matching Subsets) | | | | |
| Sampled at WMT | | | | |
| WMT22 | 8/12 | 8/12 | 9/12 | 8/12 |
| Simulation | | | | |
| -best | 7/12 | 6/12 | 8/12 | 5/12 |
| -median | 10/12 | 8/12 | 9/12 | 8/12 |
| -worst | 11/12 | 10/12 | 10/12 | 10/12 |
| Document Budgeted Snippet (Matching Subsets) | | | | |
| -best | 4/12 | 7/12 | 6/12 | 6/12 |
| -median | 7/12 | 8/12 | 9/12 | 8/12 |
| -worst | 10/12 | 11/12 | 11/12 | 9/12 |

Table 3: Effect of sampling methods comparing translation directions with document fixed snippet sampling approach to document budgeted snippet sampling.

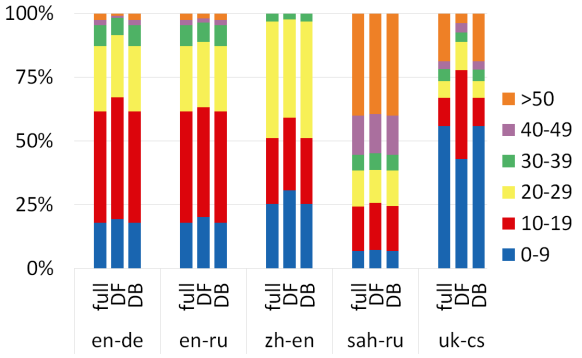


Figure 1: Proportion of data from documents with different document lengths in the full test set and the subset sampled by document fixed snippet and document budgeted snippet sampling.

| lp. | sam. | 0-9 | 10+ | 20+ | 30+ | 40+ | 50+ |
|---|---|---|---|---|---|---|---|
| en-de | full | 18% | 44% | 26% | 8% | 2% | 3% |
| | DF | 19% | 48% | 24% | 7% | 1% | 1% |
| | DB | 18% | 44% | 26% | 8% | 2% | 3% |
| en-ru | full | 18% | 44% | 26% | 8% | 2% | 3% |
| | DF | 20% | 43% | 26% | 8% | 2% | 2% |
| | DB | 18% | 44% | 26% | 8% | 2% | 3% |
| zh-en | full | 25% | 26% | 46% | 3% | — | — |
| | DF | 31% | 28% | 39% | 2% | — | — |
| | DB | 25% | 26% | 46% | 3% | — | — |
| sah-ru | full | 7% | 18% | 14% | 6% | 15% | 40% |
| | DF | 7% | 19% | 13% | 7% | 15% | 40% |
| | DB | 7% | 18% | 14% | 6% | 15% | 40% |
| uk-cs | full | 56% | 11% | 7% | 5% | 3% | 19% |
| | DF | 43% | 35% | 11% | 4% | 4% | 4% |
| | DB | 59% | 11% | 7% | 5% | 3% | 19% |

Table 4: Percentage of data from documents with different document lengths in the full test set and the subset sampled by document fixed snippet and document budgeted snippet sampling.

man's ranking correlation to present the distortions in rankings because the number of systems for each language pair ranged from 9 to 22 which does not meet the minimum number of samples needed for Spearman's ranking correlation analysis at significance level 0.05 (Bonett and Wright, 2000; May and Looney, 2020).

### 3.3 Results and Discussion

Tables 1, 2 and 3 show the number of language pairs where the ranking changed (for at least one pair of systems) between the sampled subset and the full test set; lower is better. We compare the WMT segment-level (SL), whole document (WD) and document fixed snippet (DF) sampling methods respectively against the best, median, and worst runs of document budgeted snippet (DB) sampling, and note whether the WMT methods used matching or non-matching subsets.

Comparing SL against DB in Table 1, we see that the number of inconsistent system rankings for the median run of DB is similar to that of SL. Thus, the DB approach performs similarly to SL, while having the advantage of document context. Läubli et al. (2018) and Toral et al. (2018) show that document context is necessary for annotation quality and consistency. As MT moves towards document-level translation from sentence-level, it is essential for human evaluation to also have the capability to evaluate with document context to support future research on MT. For this reason, even though our proposed DB sampling method only provides similar stability as the SL sampling method in table 1, we argue that the advantage of having document context makes DB more suitable for human evaluation of MT (Castilho (2021) makes a similar argument about tradeoffs). In Table 2, we see clearly that DB produces more consistent system rankings than WD, with the worst run of DB still having fewer inconsistent system rankings than WD. Comparing DF against DB in Table 3, we see that the number of inconsistent system rankings for the median run of DB is similar to that of DF. However, Figure 1 and Table 4 show that DF consistently oversampled segments from short documents and undersampled from long documents, relative to the proportion of the test set that they make up; DB is designed to better match the full test set distribution. It is worth additional examination to determine the consequences of a potential tradeoff between better representation of document length distributions and topic/vocabulary representation.

Beyond noting the concern about the sampled data not being representative of the full test set, our simulation method cannot demonstrate all forms of system ranking instability caused by bias in document length because current automatic MT evaluation metrics do not consider document-level quality. As MT research moves toward the document level, document length distribution in evaluation data will be increasingly important. Sampling bias in evaluation towards shorter documents may result in system rankings not able to accurately reflect system performance in translating long documents.

Let us recall that the reason we do sampling at all is because we want to understand how systems perform on the *full* test set, but do not have the budget to collect enough annotations. Coverage is a key factor in system ranking consistency, and under tight budgetary constraints it may not always be possible to mitigate instability simply by modifying the sampling method. If coverage is too low, it may be worth considering non-random alternatives, such as determining whether there are portions of the test set that are actually of greater importance to the MT use case, and selecting for those. But assuming the sampling case, we emphasize the importance of minimizing inconsistency and sources of errors in as much as possible in as many parts of the evaluation setup as possible, to prevent compounding effects on the final rankings.

We made retrospective comparisons, but there is no reason we cannot sample and compare rankings against the full test set using automatic metrics *before* performing human annotation. This could enable us to select a sample that has the smallest level of inconsistency with the full test set, rather than hoping for median performance. The risk of this is if this biases the subset due to the choice of metric (or if metrics perform poorly on the language pair); in future work we plan to examine whether such metric-guided sampling reduces inconsistency with human annotation.

## 4 Conclusions

We examine three different approaches that WMT has used for sampling segments from test sets for human judgment, performing simulations using automatic metrics in place of human annotations. This allows us to examine a large range of scenarios at low cost, with the risk that it may not be fully representative of human judgment distributions. We demonstrate in simulation that a document budgeted snippet sampling approach finds a balance between providing document context, representation (i.e., better representing the distribution of document lengths), and ranking stability. Additionally, we use this analysis to highlight problems and challenges in comparing past human annotation approaches. In particular, large and small variations in annotation procedures are often conflated and collapsed into overly-simplified descriptions that obscure the ways in which they differ from one another; we attempt to untangle some of these. We urge researchers to take care in examining – in isolation and in combination – the effects that various design decisions have on results, in order to build annotation approaches that remove as many sources of error as possible.

## Limitations

The main limitation of this work is our use of automatic metrics rather than human evaluation. First, the score distribution produced by a metric is not guaranteed to be similar to one produced by human annotators, which could influence results. Secondly, the metrics we examined do not incorporate context. Motivated by evidence that document-level (or contextual) information is becoming necessary to distinguish between human translations and high quality machine translation (Läubli et al., 2018; Toral et al., 2018), recent WMT evaluations have incorporated context. Since the human annotations are influenced by the context in which they appear and the automatic metrics are not (i.e., given an identical segment in two different contexts, the automatic metric will score them identically while a human annotator may not), additional study may be necessary to answer questions such as whether additional preceding source context should be displayed to annotators (as suggested in Castilho et al. (2020)), to determine how much additional time reading this context would take (which may influence the annotation budget), or to determine whether human annotator behavior may differ based on where in a document the snippet comes from. We also do not directly address issues such as the best interfaces for human annotation; a problem that is mostly orthogonal to the question of what data should be annotated.

In this work, we also follow the approach in the WMT Metrics shared task of treating the scores assigned to systems (in our case by automatic metrics rather than human annotators) as full rankings of systems, rather than as clusters of systems. In practice, this may mean that statistically insignificant differences between systems are considered on par with statistically significant ones when we examine reorderings that occur based on different sampling procedures. While this is a major concern in human annotation (where there is also an effort to handle annotator variation, a separate source of instability), it is less of a concern in this setting where the annotation is guaranteed to be consistent.

One additional limitation to our proposed future work of using metrics as a pre-sampling approach is that they may not perform equally well across all languages. See Appendix A for the list of language pairs on which these experiments were performed.

## Ethics/Impact Statement

This work, while it uses automatic metrics rather than human judgments to demonstrate theory, is focused on sampling methods *for* human evaluation of machine translation. Future work should examine whether human evaluation and distributions of human annotations do follow the same patterns we observed across automatic metrics in this work. A risk we have observed in failing to do adequate theoretical analysis of annotation setups is that the blame for inconsistency is sometimes shifted to the human annotators themselves, when in fact there may be more that those setting up the annotation schema ought to do to account for various other sources of inconsistency introduced into the process. Thus, we do think it is important and valuable to do additional (controlled) experiments on the approaches we have examined with human annotations, to determine whether there are user interface, context, or other issues that may present themselves in human annotation but not in automatic evaluation.

## Acknowledgements

We would like to thank the anonymous reviewers for the constructive comments on our work.

## 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, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Douglas G Bonett and Thomas A Wright. 2000. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65:23–28.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

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

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

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi,

George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68, Abu Dhabi. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. 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, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

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

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

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.

Rebecca Knowles and Chi-kiu Lo. 2022. Test set sampling affects system rankings: Expanded human evaluation of wmt20 english-inuktitut systems. In *Proceedings of the Seventh Conference on Machine Translation*, pages 140–153, Abu Dhabi. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, Maja Popović, and Mariya Shmatova. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45, Abu Dhabi. Association for Computational Linguistics.

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, New York City. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

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.

Justine O May and Stephen W Looney. 2020. Sample size charts for spearman and kendall coefficients. *Journal of biometrics & biostatistics*, 11(2):1–7.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pages 6905–6916. PMLR.

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.

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.

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, Brussels, Belgium. Association for Computational Linguistics.

## A Additional Details

In this appendix we provide average (mean) test set coverage for systems across language pairs from WMT 2019-2022.

| **Language Code** | **Language Name** |
|---|---|
| cs | Czech |
| de | German |
| en | English |
| fi | Finnish |
| fr | French |
| gu | Gujarati |
| ha | Hausa |
| hr | Croatian |
| is | Icelandic |
| iu | Inuktitut |
| ja | Japanese |
| kk | Kazakh |
| km | Khmer |
| liv | Livonian |
| lt | Lithuanian |
| pl | Polish |
| ps | Pashto |
| ru | Russian |
| sah | Yakut |
| ta | Tamil |
| uk | Ukrainian |
| zh | Chinese |

Table 5: Language codes.

| year | dir. | coverage |
|---|---|---|
| WMT19 | de-cs | 76.93% |
| WMT19 | de-fr | 35.81% |
| WMT19 | fi-en | 79.21% |
| WMT19 | fr-de | 23.52% |
| WMT19 | ru-en | 84.83% |
| WMT19 | zh-en | 67.33% |
| WMT20 | iu-en | 37.02% |
| WMT20 | km-en | 41.00% |
| WMT20 | ps-en | 43.75% |

Table 6: Average coverage per system of human annotation for the WMT19-20 test sets that used segment-level (SL) sampling method

| year | dir. | coverage |
|---|---|---|
| WTM19 | de-en | 99.84% |
| WTM19 | en-de | 86.80% |
| WTM19 | en-fi | 67.10% |
| WTM19 | en-gu | 63.08% |
| WTM19 | en-kk | 82.09% |
| WTM19 | en-lt | 74.82% |
| WTM19 | en-ru | 72.69% |
| WTM19 | en-zh | 81.41% |
| WMT20 | cs-en | 94.53% |
| WMT20 | de-en | 92.00% |
| WMT20 | en-cs | 85.67% |
| WMT20 | en-de | 47.48% |
| WMT20 | en-iu | 30.54% |
| WMT20 | en-ja | 87.07% |
| WMT20 | en-pl | 75.53% |
| WMT20 | en-ru | 62.32% |
| WMT20 | en-ta | 52.61% |
| WMT20 | en-zh | 86.87% |
| WMT20 | ja-en | 90.03% |
| WMT20 | pl-en | 84.36% |
| WMT20 | ru-en | 91.01% |
| WMT20 | ta-en | 54.29% |
| WMT20 | zh-en | 86.02% |
| WMT21 | cs-en | 90.53% |
| WMT21 | de-en | 96.89% |
| WMT21 | de-fr | 91.76% |
| WMT21 | en-ha | 77.23% |
| WMT21 | en-is | 98.53% |
| WMT21 | en-ja | 99.04% |
| WMT21 | en-ru | 93.27% |
| WMT21 | fr-de | 77.93% |
| WMT21 | ha-en | 94.36% |
| WMT21 | is-en | 89.03% |
| WMT21 | ja-en | 90.06% |
| WMT21 | ru-en | 89.26% |
| WMT21 | zh-en | 80.58% |
| WMT22 | cs-en | 92.13% |
| WMT22 | de-en | 86.71% |
| WMT22 | ja-en | 82.98% |
| WMT22 | liv-en | 70.48% |
| WMT22 | ru-en | 93.13% |
| WMT22 | uk-en | 90.68% |
| WMT22 | zh-en | 85.42% |

Table 7: Average coverage per system of human annotation for the WMT19-22 test sets that used whole document (WD) sampling method

| year | dir. | coverage |
|---|---|---|
| WMT22 | cs-uk | 60.07% |
| WMT22 | en-cs | 59.16% |
| WMT22 | en-de | 59.89% |
| WMT22 | en-hr | 69.26% |
| WMT22 | en-ja | 57.98% |
| WMT22 | en-liv | 69.90% |
| WMT22 | en-ru | 58.71% |
| WMT22 | en-uk | 59.35% |
| WMT22 | en-zh | 57.98% |
| WMT22 | sah-ru | 95.55% |
| WMT22 | uk-cs | 12.49% |
| WMT22 | zh-en | 59.63% |

Table 8: Average coverage per system of human annotation for the WMT22 test sets that used document fixed snippet (DF) sampling method

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Limitations section, 3*

☑ A2. Did you discuss any potential risks of your work?
*3*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Abstract, Introduction (1)*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*Used them, 3.1, Appendix A*

☑ B1. Did you cite the creators of artifacts you used?
*3.1, Appendix A*

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*We did not release artifacts, and readers can check the license/terms for the used artifacts by following the citations*

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Implicitly, 3.1*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Not applicable. Did not collect additional data*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Appendix*

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Not applicable. Left blank.*

### C ☑ Did you run computational experiments?

*3*

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Did not perform model training, simply sampled data subsets and computed averages. Was not run on GPU.*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*3.1*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*3*

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*No response.*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*No response.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*No response.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*No response.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*No response.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*No response.*