Test Set Sampling Affects System Rankings: Expanded Human Evaluation of WMT20 English–Inuktitut Systems

Rebecca Knowles and **Chi-kiu Lo**
National Research Council Canada (NRC-CNRC)
{rebecca.knowles,chikiu.lo}@nrc-cnrc.gc.ca

Abstract

We present a collection of expanded human annotations of the WMT20 English–Inuktitut machine translation shared task, covering the Nunavut Hansard portion of the dataset. Additionally, we recompute News rankings to take into account the completed set of human annotations and certain irregularities in the annotation task construction. We show the effect of these changes on the downstream task of the evaluation of automatic metrics. Finally, we demonstrate that character-level metrics correlate well with human judgments for the task of automatically evaluating translation into this polysynthetic language.

1 Introduction

Translation between Inuktitut[1] and English was featured as part of the 2020 News Translation task at WMT (Barrault et al., 2020). The English–Inuktitut machine translation system rankings published in Barrault et al. (2020) were incomplete, due to the delay in the 2020 annotation campaign and because they only cover the out-of-domain portion of the test set. In this work, we present:

- an expanded dataset of human annotations that covers the in-domain portion of the test set,[2]
- an analysis of both the existing (out-of-domain) human annotations and the new (in-domain) annotations,
- revised system rankings based on the new annotations and controlling for irregularities in the original data collection process,
- and correlations of automatic MT evaluation metrics with the revised system rankings and the newly collected human annotations.

[1]We use the term *Inuktitut* here because the website of the Legislative Assembly of Nunavut lists Inuktitut, Inuinnaqtun, and English (along with French) as the languages spoken in the House (https://assembly.nu.ca/faq#n113) and the version of the Hansard released as training data is the Inuktitut version, written in syllabics. See Appendix A.

[2]Dataset and code: https://github.com/nrc-cnrc/Reranking-WMT20-IKU

Dataset	Segments
Hansard-A (H-A)	11404
Hansard-B (H-B)	7801
All Hansard	19205
WMT20-DA (N-1)	8000
WMT20-DA2 (N-2)	12002
WMT20-DACrowd (N-C)	9728
All News	29730
Total	48935

Table 1: Number of segments annotated in each dataset, without any data filtering. We use names corresponding to the dataset files, with short forms in parentheses.

Our aim with this work is to release a broader set of human annotations, for continued research on translation between English and Inuktitut, as well as to demonstrate the downstream impacts of irregularities in WMT data collection and publication. In particular, we draw attention to how errors in human annotation setup (not errors in the work of the annotators, but errors in the way that organizers constructed the annotation tasks) and the failure to account for their effects impact both the validity of the rankings themselves and the shared task on automatic metrics which relies on the rankings. In this way, the 2020 WMT task on Inuktitut serves as a case study of a broader issue in the field. We provide suggestions for how to account for irregularities in existing annotation task construction and release the code and data to replicate our results.

2 Data

The test set for the WMT 2020 English–Inuktitut shared task consisted of data from the Nunavut Hansard as well as from Nunatsiaq News, collected and shared with permission. The News test data consisted of documents collected from Nunatsiaq News between September and November, 2020, as was standard for the shared task. The Hansard data in the test set, however, was collected from earlier dates. The main training data available for building constrained systems was the Nunavut Hansard 3.0

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 140 - 153

	H-A	H-B	N-1	N-2	N-C	*Total*
A	-	-	2800	-	-	*2800*
B	-	-	-	3200	-	*3200*
C	-	3801	-	200	-	*4001*
D	5600	4000	4600	2000	-	*16200*
E	-	-	600	6602	-	*7202*
F	2403	-	-	-	-	*2403*
G	3401	-	-	-	-	*3401*
Un.	-	-	-	-	9728	*9728*

Table 2: Number of segments annotated by annotator (anonymous annotator ID shown in the first column, with Un. representing all unknown annotators who annotated the crowd data) and dataset.

(Joanis et al., 2020), consisting of aligned proceedings of the Legislative Assembly of Nunavut. As a consequence, the Nunavut Hansard test data could be considered "in-domain", while the News data was "out-of-domain" (the development data was similarly divided between the two domains).

The annotators who did the work of human evaluation discussed here were fluent language experts at the Pirurvik Centre,[3] paid at professional rates. All the human annotations were collected in the segment rating with document context (SR+DC) style of direct assessment (DA; Graham et al. 2013, 2014, 2016) using the Appraise interface (see Barrault et al. 2020 for additional interface details). For each segment, annotators viewed the source sentence and a candidate translation, and scored the translation on a pseudo-continuous sliding scale from 0-100. These segments were displayed in document context, and annotators also provided document scores (we omit those from this work).

Each News story had a unique document ID, but the Hansard data was treated as a single document containing 1566 lines. This had two main consequences with respect to human annotation of system outputs. The first and most obvious is that the annotations collected at WMT (on which the Findings paper's rankings (Barrault et al., 2020) were at least partially based) only included annotations of the News data (out-of-domain). The reason for this is that the code used to generate sessions of annotations in the SR+DC DA human annotation task structures pooled all documents and then sampled documents "at random (without replacement) and assigned [them] to the current HIT [human intelligence task] until the current HIT comprise[d] no more than 70 segments in total"; since the Hansard data was treated as one document with more than 70 segments, it was never sampled.[4] The second is that the News documents were longer on average than News documents for other language pairs. English–Inuktitut News documents ranged from 12 to 137 lines in length, with a mean of 39.0 (standard deviation 20.5) and a median of 36. Documents for other language pairs that were evaluated in the SR+DC format ranged from 2 to 32 lines in length, with a mean of 12.0 lines (standard deviation 6.2) and a median of 11. As a consequence, each annotation session of English–Inuktitut News data is less likely to contain documents translated by the full set of submitted systems (12 submitted systems and human reference), which has the potential to cause problems when scores are normalized per-annotation session (Knowles, 2021).[5] Additionally, a portion of the source and reference data News segments contained spurious quotation marks (see Appendix B). We now discuss the News and Hansard annotation processes.

2.1 News Annotations

There are several other noteworthy issues about the News data collection. As shown in Table 1, there are three direct assessment datasets collected at WMT that contain News-only annotations of English–Inuktitut translations. The first, N-1, consists of 8000 segments and does not contain any annotations of reference segments. The second, N-2, contains 12002 segments and *does* contain annotations of reference segments. Both of these were annotated by fluent language experts at the Pirurvik Centre. There is a third set of data, N-C, which was annotated by other annotators (the Findings paper does not clarify who those annotators were, so we will focus our analysis on the first two datasets, known to be collected through Pirurvik). The fact that one set of data was collected with reference segments included and the other was not

[3] https://www.pirurvik.ca/

[4] Note that we will use HIT and annotation session interchangeably in this paper. An annotation session that received a single ID and thus was used as the basic chunk of data for computing z-scores typically (but not always) consisted of two HITs, each containing 100 segments. However, the way the data is released, the two HITs are not distinguishable from one another, hence our reference instead to the annotation session. Note that an individual annotator may have completed many such sessions.

[5] The use of z-score carries with it an implicit assumption that the annotation session, HIT, or set of data annotated by one annotator is representative of the whole. The raw scores for human references tended to be near-perfect, while one system's scores were zero, meaning that whether an individual annotation session contained one, both, or neither of these would unduly influence the z-score computation.

also has the potential to cause problems in generating system rankings. Because system rankings from SR+DC run through Appraise are typically calculated based on z-scores computed at the annotation session level, and because these sessions are *not* representative of the distribution of systems, they will be erroneously standardizing out real differences in quality. Even if they did compute z-scores over annotators, we can see in Table 2 that not all annotators completed annotation sessions in each dataset, meaning that some annotated reference segments and some did not; again, this means that it is inappropriate to calculate z-scores in the standard WMT fashion (even over annotators instead of over sessions). The data collection also contained quality assurance segments (these are called "BAD" segments, and quality assurance is described in more detail in Appendix C).

The system submitted under the name zlabs-nlp (no corresponding paper submitted) consisted of the exact source (English) data, but was nevertheless included in the annotation tasks. The annotators from Pirurvik received instructions to give a score of 0 to output that was not in the target language (i.e., Inuktitut) and this is reflected in their scores (almost all 0 for zlabs-nlp segments),[6] while the scores for the Crowd annotation set are much more wide-ranging (indicating that those annotators may not have received the same instructions).

2.2 Hansard Annotations

Following the completion of WMT 2020, we collected annotations of the Nunavut Hansard portion of the test set. Like the News annotations, fluent Inuktitut-language experts from Pirurvik Centre performed these segment rating with document context (SR+DC) annotations using the Appraise interface; with the help of the shared task organizers, we collected data using the same web interface as was used for the News data, allowing us to keep that portion of the annotation process consistent.

The data was processed and collected with the following noteworthy changes.[7] First, data from zlabs-nlp (exact copies of the source text) were omitted from annotation, as those scores are not representative of translation. Second, the Hansard was manually divided into pseudo-documents, ranging in length from 8 lines to 26 lines, with an average of 14.6 (standard deviation 3.7) and median 15. This is closer to the average document length for other language pairs, and enables annotation sessions to contain a more diverse set of system/document pairs. Third, in this set of annotations, references – and all systems – were more evenly distributed across annotators, improving validity of the z-score assumptions (Knowles, 2021). Finally, as shown in Table 1, the Hansard annotations were split into two parts. Wishing to ensure that all systems were annotated on consistent sets of documents, but unsure as to whether annotator time and budget would cover the full Hansard test set, we first randomly split the set of documents in two, and then generated annotation sessions by sampling from one half (Hansard-A) or the other (Hansard-B). Fortunately, annotators completed all sessions. We did not include quality assurance segments in this task, as all annotators were known to be qualified (see Appendix C).

2.3 Systems

Twelve systems were submitted to the English–Inuktitut task. In alphabetical order by team name, they were: CUNI-Transfer (Kocmi, 2020), Facebook_AI (Chen et al., 2020), Groningen (Roest et al., 2020), Helsinki (Scherrer et al., 2020), NICT_Kyoto (no corresponding paper),[8] NRC (Knowles et al., 2020), OPPO (Shi et al., 2020), SRPOL (Krubiński et al., 2020), MultiLingual_Engine_Ubiqus (Hernandez and Nguyen, 2020), UEDIN (Bawden et al., 2020), UQAM_TanLe (no corresponding paper), and zlabs-nlp (no corresponding paper). Of these twelve, Barrault et al. (2020) listed MultiLingual_Engine_Ubiqus and UQAM_TanLe as unconstrained entries (meaning that they chose to use additional data outside of those provided for the constrained version of the shared task). All systems for which we have a description used Transformer models (Vaswani et al., 2017).

3 Approaches to Rankings

In this work, we will generate two sets of rankings: system rankings over the Hansard data and system rankings over the News data. While there would be reason to desire a single ranking that covers both in-domain (Hansard) and out-of-domain

[6] The 7 scores of 1 may be simply due to slider operation.

[7] This data collection was completed prior to the publication of Knowles (2021), and as such only addresses a portion of the concerns raised in that paper. We seek to address other concerns from that paper in our analysis of the data.

[8] The Findings paper cites Marie et al. (2020) for NICT_Kyoto, but that paper does not describe an English–Inuktitut MT system.

(News), that would raise the question of how to balance the two, and would also be a challenge to produce given the differences in the data collection processes. Having two rankings also highlights differences in performance across those domains.

The Hansard rankings come from the annotations that will be released alongside this paper, while the News rankings are a reranking based on the data collected at the WMT shared task. Here we discuss how the rankings computed for this paper differ from those produced at WMT. A partial description of the WMT rankings can be found in Barrault et al. (2020). The main issues we try to address in our new rankings are those raised in Knowles (2021) around the instability of rankings, particularly when the annotation sessions contain distributional issues that make the usual z-score computation inappropriate. We attempt to handle these issues both in proactive ways (through modifications to the "document" lengths and system distributions in the setup of the annotation of Hansard data) and in reparative ways (when we make use of the existing WMT News annotations).

3.1 Hansard Ranking Approach

Ave.	Ave.z	System
89.9	0.249	SRPOL
87.5	0.201	Groningen
88.6	0.192	NICT_Kyoto
88.8	0.170	NRC
88.1	0.160	Human-A
87.1	0.133	CUNI-Transfer
85.9	0.120	Facebook_AI
85.6	0.046	UEDIN
83.6	-0.055	Helsinki
78.0	-0.127	MultiLingual_Engine_Ubiqus
76.5	-0.360	UQAM_TanLe
65.6	-0.789	OPPO

Table 3: Hansard ranking, computed using the standard WMT approach. Unconstrained systems in grey. Horizontal lines separate significance clusters.

For the Hansard rankings (Table 3), we compute them as follows. We have a mapping between annotators and annotation sessions, so for each annotator, we collect all of the data from all of their annotation sessions. Given one annotator's full set of annotations, we compute the mean m_a and standard deviation s_a (where a is the annotator). These are then used to compute the z-scores for every segment that they annotated. Given a raw score x produced by annotator a, its z-score is:

$$z = \frac{x - m_a}{s_a} \quad (1)$$

After z-scores have been computed for all segments annotated by all annotators, system scores can be computed. The first step is to average any instances of scores that share the same system ID, the same document ID, and the same sentence ID (regardless of whether they are annotated by the same or different annotators). Then, all segments produced by a particular system are averaged into the final system score. These last two steps are performed on both raw scores and z-scores, but the ranking is computed using z-scores. Clusters of systems are indicated by horizontal lines in the ranking (Tables 3 and 4), with such a horizontal line drawn below a system if and only if its z-scores are significantly better than all systems ranked below it according to a Wilcoxon ranked sum test ($p < 0.05$). The differences between this and the standard WMT data collection are the choice to compute z-scores over annotators rather than over annotator sessions and the fact that we did not collect any "BAD" quality assurance annotations.

3.2 News Ranking Approach

Ave.	Ave.z	System	Findings Ranking
90.3	0.652	Human-A	(1-2, 90.5, 0.574)
76.4	0.219	CUNI-Transfer	(3-9, 77.4, 0.409)
77.7	0.102	NICT_Kyoto	(3-9, 79.2, 0.364)
71.6	0.096	NRC	(3-9, 71.9 0.369)
76.2	0.053	Ubiqus	(1-2, 75.3, 0.425)
74.1	0.041	Helsinki	(3-9, 75.2, 0.296)
73.6	0.025	Facebook_AI	(3-9, 74.6, 0.368)
72.7	0.012	SRPOL	(3-9, 72.8, 0.282)
72.8	-0.052	Groningen	(3-9, 71.6, 0.339)
67.6	-0.305	UQAM_TanLe	(10-11, 68.9, 0.084)
65.0	-0.427	UEDIN	(10-11, 66.4, 0.081)
46.8	-1.223	OPPO	(12, 48.2, -0.384)
0.0	-3.181	zlabs-nlp	(not shown)

Table 4: News Rankings, with mean and standard deviation for z-score computed using only SRPOL (all annotators scored output from that system). The last column shows the systems' original rankings in the 2020 Findings paper: cluster range, raw average, and z-average.

For the News task (Table 4), we had the N-1 and N-2 datasets along with a mapping between annotators and annotation sessions. Starting from this, we modified the ranking computation process to attempt to account for the known concerns with the dataset. We were unable to replicate the Findings rankings (Barrault et al., 2020), nor the listed number of annotations for Inuktitut from the data released.[9] The Findings rankings may have been

[9] https://www.statmt.org/wmt20/results.html

computed from earlier, incomplete data.

Due to the high average document length (39.0 lines) and the extreme range of system quality (from human reference near 100 to zlabs at 0), we cannot expect annotation sessions to be comparable to one another and certainly not representative of the whole test data and systems. For this reason, it is already not appropriate to compute z-scores at the annotation session level. Additionally, it is not appropriate to compare z-scores that are computed in the standard way between the N-1 and N-2 datasets, since the former does not contain human references while the latter does. Adding to the challenges, not all annotators completed annotation sessions in both of the datasets, and not all annotators annotated data from all systems (or across systems in the same proportions). Thus, simply switching to the annotator-level z-score computation does not solve the problem. For this reason, we chose to compute the mean and standard deviation for each annotator based only on the SRPOL system segments that they had annotated. SRPOL and CUNI were at the intersection of all annotators' sets of annotated systems, but in different ratios, so we selected SRPOL because the annotator who had annotated the smallest number of segments had annotated more SRPOL than CUNI segments. We do not include "BAD" segments in the z-score calculations (as different annotators had different proportions of quality assurance data) and we also do not eliminate any data based on quality assurance measures. This does not guarantee that this is a perfectly fair comparison, as the specific documents and segments annotated are not consistent across annotators, but it does limit the influence of extreme outliers on the z-score computations. We then use those means and standard deviations to compute z-scores for all data across all systems.

In conjunction with these justifications, we note the following as additional support for our chosen approach to ranking the News data. The stated goal of using z-scores (rather than raw scores) in the official ranking is "to iron out differences in scoring strategies of distinct human assessors" (Barrault et al., 2020). If we had perfectly consistent annotators and were computing z-scores in such a way that they were standardizing annotator difference rather than other information in the data, z-scores and raw scores would produce matching orderings of systems. If all annotators were perfectly consistent but the z-scores did not correlate with the raw scores, then we would know that there was a problem with the z-score calculations or annotation setup. We simulate this by replacing all News human annotation scores with CHRF scores as pseudo-annotations and then calculate rankings in approximately the style of WMT20 by computing z-scores at the annotation session level[10] and then computing them using our approach. We find z-scores and raw scores produce identical system orderings under our approach, but produce less-correlated (i.e., non-identical) orderings using the WMT20 approach. Computing means and standard deviations for CHRF scores at the annotator level (but across all systems) does improve the most extreme differences between raw and z-scores, but the complete ordering is still not as well-correlated as with our new approach. While this does not guarantee that our approach fully solves the problem, it does demonstrate that our approach does not introduce the same error as the WMT20 approach.

If we were to use only SRPOL data for computing the annotator means and standard deviations for the Hansard ranking, we would obtain the same ordering of systems that we obtained via the approach described in Section 3.1 (though of course with different z-scores), with the only difference being that using SRPOL only would put UEdin and Helsinki in the same significance cluster.

4 Rankings

We observe several similarities and differences between the Hansard (Table 3) and News (Table 4) rankings. As expected, the authentic Human translations consistently score highly (with raw scores of 90.3 for News and 88.1 for Hansard) and are in the top cluster of the rankings. In the case of News, the authentic human translations are in a cluster of their own, while in the case of the Hansard data, they are in a cluster alongside the four top-performing MT systems. The raw scores for the News rankings are consistently lower than those for the Hansard rankings (both overall and on a system-by-system basis). This reflects the fact that there is less data in the News domain and the fact that the Hansard domain is highly repetitive. We will explore both of these topics in Section 5.3.

[10]This is intended to be closer to what we believe was done at WMT20; however, the WMT20 calculation for means and standard deviations likely included "BAD" reference segments, which we must omit because we do not have the "BAD" reference text to be able to compute CHRF scores against the reference.

The system that shows the smallest gap in raw scores between Hansard and News and the greatest improvement in clustering, moving from the fifth cluster for Hansard to the second for News was Ubiqus, which saw a difference of just 1.8. In comparison, the systems with the greatest drops in rankings (UEdin from third cluster to sixth, and Groningen from one to four) saw raw score average drops between 14.7 and 20.6. The OPPO and NRC systems also saw large drops in raw average scores (18.8 and 17.2, respectively) but with smaller or no corresponding ranking drops (in the case of OPPO, it was ranked last in Hansard so no drop was possible).

5 Discussion

5.1 System Performance

Here we discuss system performance across the different test sets. Table 5 summarizes some of the features of the approaches used in different submissions, while Figure 1 visualizes our two rankings and the published Findings ranking from Barrault et al. (2020). All submitted systems for which we have information used Transformer models, implemented in a range of toolkits.

System	Toolkit	BT	Tag	News Dev
CUNI	tensor2tensor	Y	-	-
Facebook	fairseq	Y	Y	75%
Groningen	Marian	Y	Y	76%
Helsinki	OpenNMT-py	Y	-	-
NICT				
NRC	Sockeye	Y	Y	100%
OPPO	fairseq	Y	-	-
SRPOL	Marian	Y	-	-
Ubiqus	OpenNMT-py	Y	-	-
UEdin	Marian	Y	-	-
UQAM				
zlabs				

Table 5: Table summarizing system features (where known), including toolkit used, use of backtranslation (BT), use of tags for domain and/or backtranslation (Tag), and whether News development data was used in training. For systems without a corresponding system description, unknown information is left blank. Unconstrained systems are marked in grey.

Two systems participated as "unconstrained" systems (incorporating additional data), with differing levels of success. Multilingual_Engine_Ubiqus moves from the bottom half of the systems when ranked on Hansard to the top half when ranked on News, and their system incorporated additional data from news and magazine domains. This may

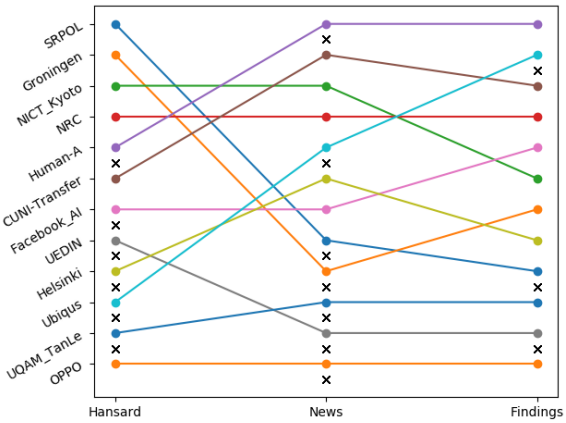

Figure 1: Summary of differences in clusterings and rankings. Black x marks indicate the demarcation between clusters and the systems are listed from best performing (top) to worst performance (bottom) across our Hansard ranking, our News ranking, and the ranking from the Findings paper (which used only News data).

account for some of the performance improvement they observed, though we cannot say with certainty if this is the main or only factor. On the other end, UQAM's system was also unconstrained but did not perform as well on News data. While examining the annotation data, we observed that despite its relatively low ranking, the UQAM system had a high number of segments with perfect automatic metric scores (CHRF of 100.0), meaning they were identical to the reference. Upon closer inspection, we found that 24.6% (295 of 1201) of UQAM segments annotated and labeled TGT (target) in the human annotation were identical to the reference. This compares to just 1.2% (183 of 14772) of all other systems' TGT annotated segments (excluding Human, which is itself the reference). The UQAM segments that were identical to the reference received very high scores, in line with the general trend for human translations. While there was not a paper submitted with the UQAM submission, the fact that it was marked as unconstrained suggests that their approach included additional data collection, and it appears that this included some of the test data.

All systems that had corresponding papers incorporated backtranslation. Three systems incorporated training on News development data (in various quantities) and these same three used tags to indicate domain and backtranslation. While this may have benefited the systems that incorporated it, it's clear that it was neither necessary nor sufficient to guarantee that a system placed in the top cluster.

Other techniques that systems used included pre-training, monolingual tasks, BPE dropout, ensembling, transfer learning, and more. There remains work to be done to identify which approaches produce the most positive impacts on translation quality. As it stands, a major challenge is that the development of the systems relied on automatic metrics, without knowing for sure which automatic metrics might be best suited to this language pair (several papers note the use of character-level metrics due to their prior results on morphologically complex languages). In Section 6 we will discuss the correlation between automatic metric scores and the human annotation results, in the hopes that this will be useful for future work on this language pair.

In addition to the rankings, we take a closer look at the performance of systems in Figures 2 and 3, which show raw human annotations averaged by document. These provide a rough visualization of system performance across different documents, as well as highlighting differences between the domains. The visualization shows both the lower overall scores assigned to the News data, as well as the greater coverage of the annotations in the Hansard data. We also see that certain documents are easy for most systems to translate, while others are consistently more difficult across systems. For example, the two documents with the highest median segment level scores, Hansard sub-documents 106 and 107, both consist of lists of names and positions, as well as standard parliamentary text about the house adjourning. Those documents with the lowest median segment scores contain longer sentences of members' speeches across varied topics like the Indspire awards, Red Seal program trades, and so on. We can also see how some systems perform relatively consistently across documents, while others exhibit more anomalous behaviour. For example we can see that the UQAM system exhibits some extremes (and the high-scored News document from 2019-11-12 is one where we note that the system output is identical to the reference, likely due to the system being unconstrained).[11]

5.2 Annotation Data Coverage

For the Hansard data collection, all systems were annotated over at least 97% of test segments, whereas for News data, coverage ranged from 47%

[11]The other systems that do see particularly high-performing documents do not have those as exact matches to the references, and in one case it is likely due to the fact that the news article is about a bill in the Legislative Assembly.

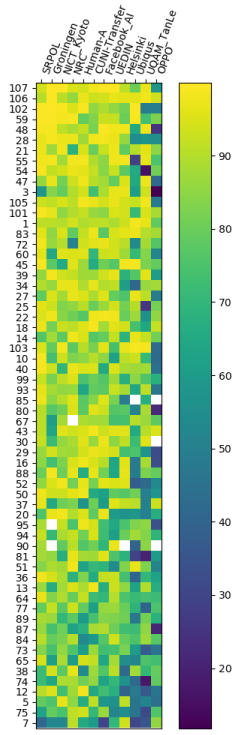

Figure 2: Raw annotation scores, averaged per document, for Hansard data. Lighter/brighter colors indicate higher scores; systems are ordered according to the rankings, while documents are ordered according to median segment score across all systems for that document. Blank spaces indicate no annotation for that document-system pair.

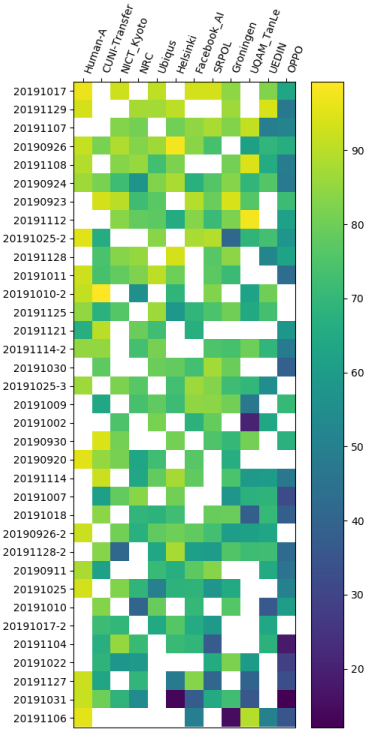

Figure 3: Raw annotation scores, averaged per document, for News data.

to 73% of test segments (with the human reference translations the least annotated).[12] This means that most Hansard "documents" were annotated for most systems (101 out of 107 had annotations for every system), while there were no News documents (out of 36) annotated for all systems. As we observe that some documents may be easier or more difficult for most systems, annotating all systems over nearly the same set of documents aims to alleviate this potential source of error.

5.3 Repetition and Novelty

Of the approximately 1.3 million sentence pairs in the Hansard training data, 59.8% of these are unique pairs, while the remainder are duplicates. The 2971 lines of the test set are all unique, and of these 15 sentence pairs were observed in the Hansard training data. If we consider only the source side, there are 155 source (English) sentences that appear in the test set that also appeared in the English side of the training data. Even though the target side of 140 of these segments is not identical to the target reference in the test set, systems still performed better on these previously observed segments than they did on segments that were previously unobserved. In fact, for all but the three lowest-performing systems, the raw scores on segments where the source had been observed in training averaged over 90.

In addition to the exact matches, the Hansard contains much boilerplate text, with small differences between what has been observed in training and the data in the test set. This includes segments like those at the start of a session, that indicate the date and time, as well as formulaic parliamentary speech (such as addressing the Speaker). All in all, the Hansard test data is more similar to the Hansard training data than the News test data is to the Hansard training data. Within each domain, there is not a strong correlation between source side similarity to training data and raw direct assessment scores, but across domains this may contribute to the differences we observe. This is also an imperfect analysis, as some systems used additional data and some incorporated development News data into their training. Nevertheless, we expect that the domain differences, compounded by the difference in data sizes, explain much of the difference in raw scores between the two domains.

[12] Adding in the Crowd annotations does not increase coverage, it simply increases the number of annotations for the sentences already annotated.

6 Automated MT evaluation

Another important area of research on English–Inuktitut machine translation is accurate automated MT evaluation metrics for a polysynthetic language. Language model based metrics usually correlate better with human judgments when evaluating translation in non-polysynthetic languages but they suffer from a training resource scarcity problem when evaluating polysynthetic languages. Character based metrics are more commonly used for evaluating translation in low resource and polysynthetic languages (Mager et al., 2021) but there is not enough study on their correlation with human judgments. A complete collection of human annotations on both domains of the English–Inuktitut test set with translation output from diverse MT systems enables further studies on automated MT evaluation metrics, with the caveat that caution should be taken with News, due to the issues in the data collection described above and in Appendix B.

6.1 Setup

We rerun the correlation analysis of the WMT20 Metrics shared task (Mathur et al., 2020b) at system level and segment level with the updated system rankings on News and the newly-collected annotations on the Hansard data. Following the Metrics shared task setup, we use `mt-metrics-eval`[13] to conduct the correlation analysis.

The correlation analysis includes all the systems, except Human-A and zlabs-nlp. Human-A is excluded because a second reference was not available for the reference based metrics to score against and zlabs-nlp is excluded because this system was not included in the WMT20 Metrics shared task test set and thus none of the participants provided scores for it.

Following the official results in WMT20 Metrics shared task, we use Pearson's coefficient to examine system level correlations of metrics with and without outlier systems. The outlier systems[14] (Mathur et al., 2020a) for the News Rankings are OPPO, UEDIN and UQAM_TanLe while those for the Hansard Rankings are OPPO and UQAM_TanLe. It is important to note that for both domains, the outlier systems are all on the lower quality side.

[13] https://github.com/google-research/mt-metrics-eval

[14] Systems that are greater than 2.5 median average deviation from the median.

	Human annotations	Findings News		News		Hansard	
	Metrics \ Systems	all	all-out	all	all-out	all	all-out
Character	characTER	0.515 (14)	0.121 (13)	0.504 (15)	-0.358 (20)	0.491 (11)	0.844 (2)
	chrF	0.336 (19)	0.091 (18)	0.355 (19)	-0.339 (17)	0.398 (14)	0.557 (12)
	chrF++	0.315 (20)	0.098 (15)	0.326 (20)	-0.323 (16)	0.344 (15)	0.566 (11)
	EED	0.483 (16)	0.122 (12)	0.495 (16)	-0.290 (15)	0.472 (12)	0.738 (6)
	YiSi-0	0.505 (15)	0.095 (16)	0.511 (14)	-0.346 (18)	0.451 (13)	0.784 (3)
Word	parbleu	0.126 (22)	0.306 (4)	0.181 (22)	-0.022 (5)	0.146 (20)	0.352 (21)
	sentBLEU	0.075 (23)	0.172 (8)	0.128 (23)	-0.152 (9)	0.048 (22)	0.503 (16)
	TER	0.357 (18)	0.083 (20)	0.441 (17)	-0.225 (12)	0.238 (18)	-0.106 (23)
Pretrn. LM	BLEURT-extended	0.762 (9)	0.155 (10)	0.759 (9)	-0.350 (19)	0.794 (7)	0.406 (19)
	COMET	0.858 (6)	0.152 (11)	0.853 (6)	-0.384 (23)	0.839 (2)	0.615 (9)
	COMET-2R	0.867 (4)	0.177 (7)	0.875 (4)	-0.152 (9)	0.725 (9)	0.735 (7)
	COMET-HTER	0.888 (3)	0.092 (17)	0.896 (3)	-0.228 (13)	0.818 (4)	0.355 (20)
	COMET-MQM	0.867 (4)	0.172 (8)	0.854 (5)	-0.368 (21)	0.825 (3)	0.463 (17)
	COMET-Rank	0.392 (17)	0.252 (5)	0.420 (18)	-0.061 (6)	0.069 (21)	0.651 (8)
	MEE	0.242 (21)	0.113 (14)	0.260 (21)	-0.285 (14)	0.219 (19)	0.579 (10)
Custom LM	YiSi-1	0.523 (13)	-0.014 (22)	0.529 (13)	-0.377 (22)	0.584 (10)	**0.852** (1)
Others	esim	0.760 (10)	0.418 (2)	0.740 (11)	-0.148 (7)	0.818 (4)	0.547 (13)
	paresim	0.760 (10)	0.418 (2)	0.740 (11)	-0.148 (7)	0.818 (4)	0.547 (13)
	prism	**0.945** (1)	0.088 (19)	**0.960** (1)	0.140 (3)	**0.974** (1)	0.775 (4)
Ref.-less	COMET-QE	0.928 (2)	**0.651** (1)	0.934 (2)	**0.534** (1)	0.298 (16)	0.237 (22)
	OpenKiwi-Bert	0.808 (8)	0.194 (6)	0.826 (8)	0.285 (2)	-0.170 (23)	0.455 (18)
	OpenKiwi-XLMR	0.680 (12)	-0.358 (23)	0.748 (10)	0.022 (4)	0.280 (17)	0.741 (5)
	YiSi-2	0.830 (7)	0.065 (21)	0.840 (7)	-0.217 (11)	0.746 (8)	0.540 (15)

Table 6: System-level Pearson's correlation of WMT20 Metrics shared task participants with z-score reported in WMT20, table 4 and 3. For WMT20 News and table 4 rankings, the outlier systems are UQAM_TanLe, UEdin and OPPO. For Hansard, the outlier systems are UQAM_TanLe and OPPO.

6.2 System-level correlation

Table 6 shows system-level Pearson's correlations of metrics with revised rankings on the News domain and new rankings on the Hansard domain.

When the outliers are included, the system-level correlations with the revised rankings are similar to those reported in the WMT20 Metrics shared task, based on the Findings rankings. However, we have several striking observations on the correlation with the revised rankings excluding the outlier systems:

- Most metrics show negative correlations with the revised rankings on the News domain.
- Reference-less metrics correlate better with revised rankings than reference based ones do.
- The rankings of the automated metrics change drastically when comparing against those obtained by correlating with the rankings in Barrault et al. (2020).
  - prism (Thompson and Post, 2020) and OpenKiwi-XLMR (Kepler et al., 2019) change from having the lowest correlation with the Findings rankings to being some of the very few metrics with a positive correlation with the revised rankings.
  - Similar changes can also be observed in YiSi-2 (Lo and Larkin, 2020) and TER (Snover et al., 2006) where they change from having the lowest correlation to the middle of the pack.
  - On the contrary, BLEURT-extended (Sellam et al., 2020), COMET (Rei et al., 2020), COMET-MQM and characTER (Wang et al., 2016) demote from the middle of the pack to having the lowest correlation with the revised rankings.

As we have established in section 3.2 that we believe the revised News rankings to be more accurate, the negative correlations with human achieved by the majority of the metrics reflect the difficulty in evaluating translation quality of low-resource polysynthetic languages for out-of-domain settings. It is important to note that the range of z-scores in the revised News rankings is [-0.052, 0.219]. It is a noticeably smaller range as compared against the range of z-scores in the Hansard rankings, which is [-0.127, 0.249]. The small variation of MT system performance in the News domain also increases the difficulty of the automated evaluation task.

Prism performs consistently well across domains with and without outliers. This is perhaps because it is one of the very few metrics that used the constrained English–Inuktitut data to train their metrics to evaluate translation quality in Inuktitut.

For the Hansard domain, it is not surprising to see YiSi-1 (Lo, 2020) correlating very well with hu-

	Metrics \ Annotations	Findings News	News	Hansard	News+Hansard
Character	characTER	0.309 (11)	0.333 (11)	0.265 (6)	0.289 (6)
	chrF	0.344 (5)	0.373 (5)	0.293 (2)	0.321 (2)
	chrF++	0.338 (6)	0.368 (6)	0.288 (3)	0.317 (4)
	EED	0.361 (3)	0.395 (3)	0.277 (4)	0.319 (3)
	YiSi-0	0.362 (2)	0.396 (2)	0.268 (5)	0.313 (5)
Word	parbleu	0.212 (14)	0.232 (15)	-0.043 (19)	0.054 (18)
	sentBLEU	0.206 (15)	0.233 (14)	-0.004 (18)	0.080 (15)
	TER	-0.071 (21)	-0.051 (21)	-0.284 (23)	-0.201 (23)
Pretrained LM	BLEURT-extended	0.359 (4)	0.387 (4)	0.226 (7)	0.283 (7)
	COMET	0.322 (9)	0.342 (9)	0.147 (11)	0.216 (9)
	COMET-2R	0.326 (8)	0.344 (8)	0.143 (12)	0.214 (11)
	COMET-HTER	0.331 (7)	0.348 (7)	0.135 (13)	0.211 (12)
	COMET-MQM	0.313 (10)	0.337 (10)	0.127 (14)	0.202 (13)
	COMET-Rank	0.297 (12)	0.312 (12)	0.174 (10)	0.223 (8)
	MEE	-0.074 (22)	-0.054 (22)	-0.212 (22)	-0.156 (22)
Custom LM	YiSi-1	0.251 (13)	0.269 (13)	0.186 (9)	0.215 (10)
Others	esim	0.122 (17)	0.142 (17)	0.039 (15)	0.075 (16)
	paresim	0.122 (17)	0.142 (17)	0.039 (15)	0.075 (16)
	prism	**0.452** (1)	**0.475** (1)	**0.326** (1)	**0.379** (1)
Reference-less	COMET-QE	-0.040 (20)	-0.036 (20)	-0.084 (20)	-0.067 (20)
	OpenKiwi-Bert	-0.115 (23)	-0.098 (23)	-0.169 (21)	-0.143 (21)
	OpenKiwi-XLMR	0.060 (19)	0.062 (19)	0.036 (17)	0.045 (19)
	YiSi-2	0.146 (16)	0.147 (16)	0.189 (8)	0.174 (14)

Table 7: Segment-level Kendall's correlation of WMT20 Metrics shared task participants with raw scores collected in News (N-1 and N-2), Hansard (H-A and H-B) and News+Hansard.

mans when the outliers are excluded. It is because YiSi-1 is based on XLM (Lample and Conneau, 2019) trained on the constrained English–Inuktitut parallel training data in Hansard domain. Another observation is that for evaluating in-domain systems, character-based metrics, characTER and YiSi-0, correlate very well with humans. This is the first scientific evidence that character-based MT evaluation metrics are a better choice for evaluating translation quality in low-resource polysynthetic languages.

6.3 Segment-level correlation

Table 7 shows the segment-level Kendall's correlations of metrics. We observe much more consistency in metrics' correlation with humans at segment level than that at system level across domains. This is possibly due to the fact that there are more data points used for correlation analysis at the segment level than the system level. Similar to correlations at system level, prism consistently correlates the best with humans at segment level.

We see even stronger evidence here at segment level that all character-based metrics (Wang et al., 2016; Popović, 2015, 2017; Stanchev et al., 2019; Lo, 2019) correlate very well with humans for evaluating translation quality in polysynthetic languages across domains. This is a particularly important finding because these character-based metrics are resource-free. That means we now have strong confidence in using character-based metrics for evaluating translation quality in a low-resource polysynthetic language.

7 Conclusion

In this work we present additional human annotations for the Hansard portion of the WMT 2020 English–Inuktitut machine translation shared task test set. We provide new system rankings on this portion of the data and present revised rankings on the News portion. We demonstrate that these changes in rankings have downstream effects on the evaluation of automatic metrics for the shared task, and examine the difficulty of performing automatic evaluation on out-of-domain text in a polysynthetic language. When it comes to automatic metrics, we find that the top-performing system incorporated training data in the low-resource target language. However, character-level automatic metrics (which did not require training) also performed amongst the top systems, demonstrating their appropriateness for evaluating translation into Inuktitut. While additional research will be required to confirm that this finding generalizes to other polysynthetic languages, we release this expanded dataset to enable more study of automatic metrics for low-resource and polysynthetic languages.

Acknowledgements

We thank the language experts and our contacts at Pirurvik Centre for their work on the annotation tasks. We thank Roman Grundkiewicz, Tom Kocmi, and Christian Federmann for their assistance in preparing and hosting the second round of Appraise annotations. We thank Eric Joanis for his work on organizing the WMT20 task, preparing the data for the task and annotation, providing information about task details, and for his feedback. We thank Roland Kuhn for his role in organizing the shared task and for his feedback, and Darlene Stewart and Samuel Larkin for work during the WMT20 shared task that informed this work. We thank Gabriel Bernier-Colborne, Cyril Goutte, Michel Simard, Yunli Wang, Patrick Littell, our colleagues, and the anonymous reviewers for their suggestions and comments.

References

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

Rachel Bawden, Alexandra Birch, Radina Dobreva, Arturo Oncevay, Antonio Valerio Miceli Barone, and Philip Williams. 2020. The University of Edinburgh's English-Tamil and English-Inuktitut submissions to the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 92–99, Online. Association for Computational Linguistics.

Peng-Jen Chen, Ann Lee, Changhan Wang, Naman Goyal, Angela Fan, Mary Williamson, and Jiatao Gu. 2020. Facebook AI's WMT20 news translation task submission. In *Proceedings of the Fifth Conference on Machine Translation*, pages 113–125, Online. Association for Computational Linguistics.

Benoît Farley. 2009. The Uqailaut Project.

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.

François Hernandez and Vincent Nguyen. 2020. The ubiqus English-Inuktitut system for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 213–217, Online. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

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

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2020. NRC systems for the 2020 Inuktitut-English news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 156–170, Online. Association for Computational Linguistics.

Tom Kocmi. 2020. CUNI submission for the Inuktitut language in WMT news 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 171–174, Online. Association for Computational Linguistics.

Mateusz Krubiński, Marcin Chochowski, Bartłomiej Boczek, Mikołaj Koszowski, Adam Dobrowolski, Marcin Szymański, and Paweł Przybysz. 2020. Samsung R&D institute Poland submission to WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 181–190, Online. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Tan Ngoc Le and Fatiha Sadat. 2020. Low-resource NMT: an empirical study on the effect of rich morphological word segmentation on Inuktitut. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 165–172, Virtual. Association for Machine Translation in the Americas.

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.

Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 895–902, Online. Association for Computational Linguistics.

Chi-kiu Lo and Samuel Larkin. 2020. Machine translation reference-less evaluation using YiSi-2 with bilingual mappings of massive multilingual language model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 903–910, Online. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Combination of neural machine translation systems at WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 230–238, Online. Association for Computational Linguistics.

Joel Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and using an English-Inuktitut parallel corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 115–118.

Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 65–74, Ann Arbor, Michigan. 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.

Jeffrey Micher. 2017. Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106, Honolulu. Association for Computational Linguistics.

Jeffrey Micher. 2018. Using the Nunavut hansard data for experiments in morphological analysis and machine translation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 65–72, Santa Fe, New Mexico, 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.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Christian Roest, Lukas Edman, Gosse Minnema, Kevin Kelly, Jennifer Spenader, and Antonio Toral. 2020. Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 274–281, Online. Association for Computational Linguistics.

Yves Scherrer, Stig-Arne Grönroos, and Sami Virpioja. 2020. The University of Helsinki and aalto university submissions to the WMT 2020 news and low-resource translation tasks. In *Proceedings of*

the Fifth Conference on Machine Translation, pages 1129–1138, Online. Association for Computational Linguistics.

Lane Schwartz, Francis M. Tyers, Lori S. Levin, Christo Kirov, Patrick Littell, Chi-kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, Jeffrey Micher, Lonny Strunk, Han Liu, Coleman Haley, Katherine J. Zhang, Robbie Jimerson, Vasilisa Andriyanets, Aldrian Obaja Muis, Naoki Otani, Jong Hyuk Park, and Zhisong Zhang. 2020. Neural polysynthetic language modelling. *CoRR*, abs/2005.05477.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Online. Association for Computational Linguistics.

Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, and Jie Hao. 2020. OPPO's machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 282–292, 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.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520, Florence, Italy. Association for Computational Linguistics.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

A Context and Related Work

There is a dialect continuum of Inuit languages, including Inuktitut, that spans Arctic communities from Alaska to Greenland. The term Inuktut is often used to refer to parts of that dialect continuum, including Inuktitut.[15] There are two main orthographies used to write these languages: Roman orthography (Latin alphabet, *qaliujaaqpait*) and syllabics (*qaniujaaqpait*).[16] The language is morphologically complex – individual words are constructed of multiple morphemes – and a word may correspond to a whole phrase or more when translated into English.

There has been a range of computational work on Inuktitut over the past decades. This includes early work on alignment and the Nunavut Hansard (Martin et al., 2003, 2005) and the recent release of a new version of the aligned Nunavut Hansard, used as training data in this task (Joanis et al., 2020). Morphological analysis and segmentation have also been areas of interest (Farley, 2009; Micher, 2017). There is also prior work on machine translation (Micher, 2018; Schwartz et al., 2020; Joanis et al., 2020; Le and Sadat, 2020).

There has been limited to no work on human and automatic evaluation of machine translation into Inuktitut prior to this work. Prior work has shown that character-based automatic metrics demonstrate promising performance on morphologically rich languages, at least in part because they do not penalize morphological variation as much as word-level exact-match metrics do (Stanojević et al., 2015; Popović, 2016). Put another way, they award "partial credit" when a system produces some but not all of the morphemes of a word correctly. This is particularly important when translating into polysynthetic or morphologically complex languages. While our results in this paper show the promise of character-level metrics, it would be useful for future work to provide a more in-depth examination of their performance to better understand their success, perhaps with analysis at the word level and not simply the sentence level.

[15] https://tusaalanga.ca/about-Inuktut

[16] https://tusaalanga.ca/node/2505

B Quotation Marks

During the test set submission period at WMT20, it was noted that a number of segments in the test set were wrapped in ASCII quotes. This was specifically an issue with the News portion of the test set; 844 News segments exhibited this ASCII quote wrapping on source, target, or both, while just 561 of the News segments were unaffected by this. As the submission period was already underway, the task organizers made the decision not to change the test set and indicated that the annotators would be told not to take the quotation mark issues into account during their evaluation.

There remain, however, several ways that this problem may have impacted the task and its results. The first is that it may have altered the behavior of MT systems, as different systems may be more or less robust to this kind of variation in input. As we do not have access to most of the MT systems, we cannot test this. The second is that teams may have handled this differently, with some adding specialized preprocessing to deal with the wrapped quotation marks and others not, and not all system description papers indicate whether or not there was special handling of this issue. Lastly, it can have an effect on automatic metric behavior. We explore that briefly below.

If we examine just the set of segments with these spurious quotations on the target side, and compute BLEU using the segments with quotes as the reference, and identical segments but with the quotes removed as the hypothesis, we see the BLEU score drop more than 10 points (from a perfect score). Since there are so many segments with these quotation marks, we still see drop of more than 5 points when we expand to the full news portion of the test data. The impact on CHRF scores is smaller.

These spurious quotation marks, while not semantically meaningful, have varied impacts on automatic metric scores, and may have also had varied impacts on translation performance across MT systems. Unfortunately, because they make up such a large portion of the News portion of the test set, omitting them dramatically shrinks the pool of data available for computing rankings and correlations. Thus, we present this work with them included, and provide these caveats about the data.

C Quality Assurance

The quality control task used in out-of-English translation directions at WMT 2020 was "BAD reference pairs", which are segments where a short segment of a translation is randomly replaced with an equal length segment randomly selected from a different reference segment. For more details on their construction see (Barrault et al., 2020). The theory is that an annotator should score the "BAD" version of a segment lower than the original version of the same segment. If an annotator does not do so over the course of an annotation session, that session would be removed.

We note that there is a reason to not fully trust this particular approach to quality control for the News dataset. The system submitted under the name zlabs-nlp (no corresponding paper submitted) consistently received scores of 0 because it was identical to the English source. In most cases, the "BAD" references paired with zlabs-nlp segments also received scores of zero, but in a few cases they received low but non-zero scores. Unfortunately, because the text of the "BAD references" were not released by the organizers, we cannot examine this more closely, or determine whether this problem may also extend to other systems.

The quality assurance tasks typically used at WMT are included in order to exclude annotators' data from the final evaluation; in particular this would include annotators who are not adequately familiar with the language pair, who are not performing careful analyses, or who might be attempting to game a crowdsourcing task. While it may be easy to simply replace annotators for certain language pairs with very large bilingual populations, there is a much smaller number of fluent bilingual speakers of English and Inuktitut. This, combined with the very high demand for their language skills (e.g., in translation), meant that we chose to work with the Pirurvik Centre, who recruited a small number of highly-skilled fluent speakers to participate in this work. Thus, the annotators' language skills and quality of work (in different language-related tasks) are known to be high (unlike in a crowdsourcing scenario, where little information is typically known about participants).

In future work and under less tight constraints as regards annotator time and budget, we would encourage the collection of repeated annotation data. This could include repeated annotations performed by the same annotator (intra-annotator agreement) as well as repeated annotations across annotators (such as a calibration HIT that all annotators complete to examine inter-annotator agreement).