Barriers to Effective Evaluation of Simultaneous Interpretation

Shira Wein[1]*, **Te I**[2], **Colin Cherry**[2]
Juraj Juraska[2], **Dirk Padfield**[2], **Wolfgang Macherey**[2]
[1]Georgetown University, [2]Google
[1]sw1158@georgetown.edu
[2]{tei,colincherry,jjuraska,padfield,wmach}@google.com

Abstract

Simultaneous interpretation is an especially challenging form of translation because it requires converting speech from one language to another in real-time. Though prior work has relied on out-of-the-box machine translation metrics to evaluate interpretation data, we hypothesize that strategies common in high-quality human interpretations, such as summarization, may not be handled well by standard machine translation metrics. In this work, we examine both qualitatively and quantitatively four potential barriers to evaluation of interpretation: disfluency, summarization, paraphrasing, and segmentation. Our experiments reveal that, while some machine translation metrics correlate fairly well with human judgments of interpretation quality, much work is still needed to account for interpretation strategies during evaluation. As a first step to addressing this problem, we develop a fine-tuned model for interpretation evaluation, which achieves better correlation with human judgments than state-of-the-art machine translation metrics.

1 Introduction

Simultaneous interpretation is an especially difficult type of translation because it requires the system or human to convey the ideas from one language to another in real time. Due to the cognitive load and constraints on memory associated with the act of human interpretation, the number of errors increases exponentially after only minutes of interpreting (Moser-Mercer et al., 1998). To compensate for these challenges, interpreters often make use of a range of strategies, such as summarization and segmentation (He et al., 2016), to concisely provide the gist of what is being said in the source language.

Despite the prevalence of both human simultaneous interpretation and automatic interpretation models, investigations into how to effectively evaluate the quality of interpretation data are extremely limited.[1] Recent work suggests that standard automatic machine translation metrics are appropriate for interpretation, due to a correlation of select MT metrics (namely BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and METEOR (Banerjee and Lavie, 2005)) with human judgments of interpretation quality (Lu and Han, 2023) and the use of METEOR for interpreter quality assessment (Stewart et al., 2018).

Recent work has also argued that simultaneous interpretation evaluation systems should be trained and tested on interpretation data as opposed to translation data (Zhao et al., 2021). In support of this argument, Zhao et al. (2021) demonstrate that there is a sizable difference in BLEU score (13.83 points) when evaluating based on interpretation or translation data.

Given the strategies unique to human interpretation and indications in prior work as to the potential utility of machine translation (MT) metrics, our goal in this work is to investigate the applicability of both (1) interpretation data as references, and (2) existing machine translation metrics for evaluation of interpretation. We argue that the strategies that interpreters leverage to be able to perform live interpretation are critical to the task and should not be penalized by the evaluation metric.

Thus, we pose three primary questions:

1. Do human interpretations collected for other purposes have sufficient quality to be considered for use as references in evaluation?

2. Can we use existing machine translation metrics—as they are—to evaluate interpretation data?

[*] Work completed while interning at Google.

[1] The study of the evaluation of simultaneous translation **latency** is quite active. However, this paper concerns itself only with evaluating the quality (i.e. adequacy and fluency) of an interpretation, ignoring the temporal axis altogether.

Findings of the Association for Computational Linguistics: EACL 2024, pages 209–219
March 17-22, 2024 ©2024 Association for Computational Linguistics

3. Can we develop a refined automatic metric that achieves higher correlation with human judgments of interpretation quality and accounts for common features of interpretations?

To carry out these research questions, we analyze and evaluate both human interpretations and machine translations, identifying potential interpreter strategies that may degrade metric effectiveness (Section 3.4). For meta-evaluation, we conduct a human evaluation on the quality of both human interpretation and machine translation to see how those metrics correlate with human judgments (Section 4.1). We then conduct a study to assess the sensitivity of the metrics when these strategies are present in an interpretation (Section 4.2). Finally, in order to further improve the correlation with human judgments, we adapt the method from MetricX (Juraska et al., 2023) and create a fine-tuned model using our interpretation data and human annotations (Section 4.3). We demonstrate that our new metric is better at assessing interpretation quality, achieving higher correlation with human judgments, suggesting that fine-tuned neural metrics can be valuable tools for assessing interpretation.

2 Related Work

Common strategies in interpretation include segmentation, passivization, generalization, and summarization (He et al., 2016; Al-Khanji et al., 2000). Bernardini et al. (2016) also show that interpretations are consistently simpler than their translated counterparts, having lower lexical density, lower mean sentence length, and greater use of frequent words.

Regarding the use of interpretation data as references, Zhao et al. (2021) show that there is a 13.83 gap in BLEU score when evaluating simultaneous machine translation output against interpretation transcripts versus the revised text translation. The decrease in system performance when evaluating against interpretation data can also be observed in Machácek et al. (2021) and Xiong et al. (2019). The differences between how translators and interpreters translate speech is notable; still, there is no consensus on how to use automatic metrics to evaluate interpretation.

Within the realm of interpretation evaluation, Fantinuoli and Prandi (2021) adapt a framework developed for human interpreter assessment and perform a human evaluation of both interpreters and machine translation systems. They find that interpreters perform better in intelligibility than machine translation systems, but worse in terms of informativeness. Macháček et al. (2023) recommends COMET (Rei et al., 2020) as a metric for assessing automatic simultaneous speech translation, though the systems considered do not mimic interpreter strategies such as summarization.

Recent work has also perturbed machine translation data in order to investigate the sensitivity of MT evaluation metrics to different types of errors (Karpinska et al., 2022). We adapt this idea in our work to investigate the sensitivity of MT metrics to different interpretation strategies. Per the results of WMT22, MetricX and COMET are the highest ranked automatic MT evaluation metrics when ranked via agreement with human judgments of machine and human translations (Freitag et al., 2022).

A number of multilingual interpretation corpora have been developed in prior work. Shimizu et al. (2014) collect an English↔Japanese interpretation corpus and show that the most experienced interpreter achieves the highest BLEU score. Doi et al. (2021) present the NAIST dataset, which is a larger English↔Japanese interpretation corpus, and using a similar setup as Shimizu et al. (2014), show that the most experienced interpreter also has a higher BERTScore (Zhang et al., 2019). However, they point out that BERTScore fails when interpreters use a strategy like summarization. The VoxPopuli corpus includes simultaneous interpretation data of European Parliament event recordings in 24 languages (Wang et al., 2021). Zhang et al. (2021) also collect a Chinese to English interpretation corpus with three experienced interpreters. Depending on whether the interpreters' performance is based on human judgments or BLEU scores, the interpreters rank differently in terms of performance.

3 Methodology

In order to assess the presence of barriers to effectively evaluating interpretation data, we leverage comparisons between simultaneous interpretation data and machine translation data (as described in Section 3.1); we perform a human evaluation study on the interpretations and machine translation data (Section 3.2) to collect human judgments of both fluency and adequacy. We use five machine translation metrics (Section 3.3) to assess the applicability of existing metrics in evaluating interpretation data, and identify features in the interpretation data which may impact metric correlation with human judgments (Section 3.4).

3.1 Data

We use the European Parliament Translation and Interpreting corpus (EPTIC; Bernardini et al., 2016) to create three data points: (1) the reference, (2) the interpretation, and (3) an in-house machine translation. The original source data are Italian remarks, read from a pre-written script. We take as our reference the provided human English translations of the Italian script. The interpretations are real-time English simultaneous interpretations produced by expert interpreters. The machine translations were obtained by translating the provided transcriptions of the Italian source audio, using the publicly available Google Translate API.[2] The dataset consists of 67 documents. We chose to use the EPTIC dataset for our experiments because of its size and the comparatively (against similar corpora) high quality of the included simultaneous interpretations.

In order to facilitate manual analysis, we break the documents in the EPTIC remarks down to the sentence level. Splitting these documents into aligned sentence pairs is difficult due to various interpretation strategies, such as summarization, omission, and segmentation. Therefore, we first align the unsegmented interpretation with the reference sentences by minimizing word error rate (WER; Matusov et al., 2005). This automatic alignment worked well for shorter documents, but it required extensive manual corrections for about half of the documents. From the 67 documents, we obtained 590 aligned sentence triplets (with each triplet again consisting of the reference, interpretation, and machine translation).

3.2 Human Evaluation Study

We collect sentence-level judgments of the interpretations and machine translations described in Section 3.1. The machine translation and interpretation are presented to the raters side-by-side, as well as the reference. In order to mask the identity of the interpretation and limit bias in annotation, we remove minor disfluencies (e.g. 'uhm') and randomize the presentation of the data such that the side that the translation appears on is consistent. We collect judgments from 1-4 for fluency and adequacy, with adequacy evaluated in comparison to the reference. In addition, examples are given in the rater template for each choice. The judgments are collected from two fluent speakers of English and are z-normalized. For adequacy, raters were instructed that omission of non-essential or non-core content is acceptable for the "Most" grade, and disfluency and segmentation errors (e.g. words from other sentences incorrectly appended to the example) should also be ignored. Four adequacy options are presented to raters:

1. **None**: Absolutely none of the meaning of the input is represented by the output. The two texts are totally unrelated.
2. **Little**: Some of the meaning of the input is conveyed by the output, but much is missing, or a lot of extra meaning has been added.
3. **Most**: Most of the meaning of the input is conveyed by the output. Some detail or nuance may be lost, or the output might include a little extra meaning absent from the input.
4. **All**: All of the meaning and nuance of the input is conveyed by the output, with no extra meaning added.

For fluency, four choices are given:

1. **Nonsense**: Not understandable as English text.
2. **Poor**: Many or serious spelling, grammar, or other mistakes, which make the text difficult to understand or hard to read. It seems to be written by somebody who doesn't know English well.
3. **Good**: Few or minor spelling or grammar mistakes; the text is still mostly understandable and readable.
4. **Flawless**: Perfect use of English with no mistakes at all.

3.3 MT Metrics

In order to investigate the utility of existing machine translation metrics for evaluating interpretation data, we employ five machine translation metrics:

1. BLEU[3] (Papineni et al., 2002)

[2] https://translate.google.com/

[3] For BLEU scores, we use sacreBLEU (Post, 2018) version v2.3.0.

2. METEOR[4] (Banerjee and Lavie, 2005)

3. BERTscore[5] (Zhang et al., 2019)

4. MetricX[6] (Juraska et al., 2023)

5. COMET[7] (Rei et al., 2020)

BLEU and METEOR are both n-gram-based metrics that calculate the similarity between the hypothesis translation and the reference n-grams.

BERTScore computes the similarity of the candidate and reference as the sum of cosine similarities between their token embeddings.

MetricX and COMET are both neural metrics which rely on contextual language model embeddings and are fine-tuned with human assessments. While MetricX and COMET differ in their neural network architectures, both optimize regression objectives on direct assessment (DA) data and Multidimensional Quality Metrics scores (Lommel et al., 2014; Freitag et al., 2021) that have been collected by WMT over the years. However, no interpretation data has thus far been used to train these metrics.

In Section 4.3, we adopt MetricX with an mT5 XL backbone (Xue et al., 2021) for further fine-tuning with interpretation data. Our first approach uses the z-normalized human annotation scores of our interpretation data (from Section 3.2) to fine-tune the base model. Our second approach fine-tunes the base model first with WMT DA data and then with our annotations. In this way, the model first learns the translation assessment task, which is then adapted to handle interpretations.

3.4 Measuring Metric Sensitivity to Interpretation Features

To investigate how well these MT metrics accommodate the strategies interpreters use to be able to translate in real time, we compare metric scores for human interpretation of audio against the output of machine translation applied to a human transcript of the same audio. We do this by manually iterating item-by-item through every interpretation/translation pair, noting instances where the machine translation score is much higher than the interpretation score. This allows us to identify features of interpretation which may degrade their scores according to current metrics. Then, we classify the type of difference between the interpretation and MT sentences to identify common individual features that seem to be having an effect on evaluation.

Through this rigorous manual process, we identify four features of interpretation that may degrade their scores according to current metrics: (1) disfluency, (2) summarization, (3) paraphrasing, and (4) segmentation.

Though we have identified these features as potentially having an impact qualitatively on metric score, we set out to quantitatively measure the impact of each feature. To see how each feature of interpretations impacts metrics, we use automatic methods to either remove the feature from our interpretation data, or add the feature to our machine translation data, and then re-compute the metric scores. This enables us to quantify the specific impact of the feature on the metric score.

For disfluency, we use the 12-layer `small-vocab` BERT disfluency detection model from Rocholl et al. (2021) to remove disfluencies from the interpretation.

For summarization and paraphrasing, we use the instruction-tuned PaLM-2 Bison LLM (Anil et al., 2023) to perturb machine translation data, prompting the model to apply summarization or paraphrasing. We iterate over multiple prompts and manually verify the quality of the LLM output in order to ensure that we have engineered the most effective prompt for this task. Specifically, we verify that the selected prompt sufficiently maintains meaning and fluency in the summarized/paraphrased output through manual analysis. Once we selected the specific prompt ("Apply summarization to the following sentence: [sentence to be summarized]. Do not include the word summarization in the response, just output the summarized sentence."), we ran the LLM over all of the machine translation data to collect a summarized and paraphrased version of each item. The paraphrase prompt was analogous, swapping in the word 'paraphrasing' for 'summarization.'

Lastly, for segmentation, we employ document-level automatic MT metrics to evaluate the document pairs.

[4] We use the implementation of METEOR from NLTK (Bird and Klein, 2009) version 3.8.1.

[5] We re-implement the BERTScore algorithm, using the pre-trained model "BERT-Base, Multilingual Cased" from Turc et al. (2019).

[6] We use an internal implementation of sentence-level and document-level MetricX models from Juraska et al. (2023).

[7] For COMET, we use `wmt22-comet-da`.

Metric	SI	MT
BLEU	0.1811	0.3276
METEOR	0.3966	0.6226
BERTScore	0.8122	0.8812
MetricX	0.5928	0.7351
COMET	0.6809	0.7818

Table 1: Average scores for simultaneous interpretation (SI) and machine translation (MT) data on automatic machine translation metrics.

4 Results

In the subsections that follow, we address each of our research questions. Namely, in Section 4.1 we address whether human interpretations (collected for other purposes) have sufficient quality to be considered for use as references in evaluation. Then, in Section 4.2, we ascertain whether we can use existing machine translation metrics—as they are—to evaluate interpretation data. Finally, in Section 4.3, we develop a refined automatic metric which achieves higher correlation with human judgments of interpretation quality and accounts for common features of interpretations.

4.1 Evaluating Human Interpretation

To address our first research question (whether interpretations have sufficient quality to be used as references), we evaluate the interpretation data and machine translation data using the MT metrics. Then, we contrast both sets of scores to reveal any deficiencies in individual interpretations.

As shown in Table 1, all metrics score the machine translation data higher than the interpretation data. This finding is in line with previous work (Xiong et al., 2019; Zheng et al., 2020).

This observation may reflect a flaw in the metrics rather than the interpretations; therefore, we move to our human evaluation, shown in Table 2. Via our human evaluation, we find that 350 out of 590 of the interpretations are missing full adequacy/ meaning preservation, whereas this is the case for only 133 of the 590 machine translations. All human ratings are lower for the interpretation than for the MT, with adequacy being the primary issue. We also observe numerous low quality interpretations in the dataset such as the example in Table 3, calling into question whether we can use interpretations as references. In this drastic example, the interpretation has a MetricX score of 0.4691 and the MT has a MetricX score of 0.7913.

	Avg Fluency	Avg Adequacy
Interpretation	3.733	3.173
MT	3.848	3.748

Table 2: Average human evaluation scores for fluency and adequacy of the interpretation and machine translation data.

Ref: "Your collective efforts were crucial in reaching a turning point in negotiations between the European institutions on this extremely technical dossier."

MT: "Collective efforts, your collective efforts have been instrumental in reaching a breakthrough during the negotiations between the institutions on this highly technical dossier."

SI: "The collective efforts of honourable members were crucial in achieving ehm crossroads and making process in what i- progress in what is an extremely technical... issue"

Table 3: Example of a low quality interpretation found in the EPTIC dataset.

Ultimately, our findings both from the automatic metrics and our human evaluation suggest that there are issues in the interpretation data that make it unsuitable for use as a reference. Specifically, the issue of low adequacy, due to content dropping and high cognitive load, causes interpretations to be insufficiently reliable to serve as references in system evaluation. While omission and summarization are to be expected in real-time interpretation, low-quality interpretations (such as the interpretation featured in Table 3) are also present.

4.2 Suitability of MT Metrics for Interpretation

To address our second research question (should we use MT metrics to evaluate interpretations), we first ask: do metrics actually correlate well with human judgments of interpretation quality?

Table 4 shows segment-level correlation between our human judgments and the automatic metrics. We find that the correlation is low compared to previous work (e.g. Sellam et al. (2020)). By examining cases where human and automatic judgments disagree, we can easily find cases where the interpreter is doing a good job, but the metric scores are low. This suggests that metric scores are overly sensitive to features of interpretation that appear in high-quality interpretations. Through qualitative analysis, we find four features of interpretation that metrics may not be handling well (potential "metric failures"): (1) segmentation, (2) minor disfluencies, (3) summarization, and (4) paraphrasing.

Next, we quantify the sensitivity of metrics to each of these four features by using the experi-

Metric	SI Fluency	SI Adequacy	MT Fluency	MT Adequacy
BLEU	0.1321	0.3999	0.0755	0.2872
METEOR	0.0819	0.5913	0.0368	0.3746
BERTScore	0.1181	0.5985	0.0843	0.3781
MetricX	0.2290	0.6023	0.1935	0.4436
COMET	0.2397	0.6306	0.1773	0.4451

Table 4: Pearson's correlation between human judgments of fluency and adequacy for the simultaneous interpretation (SI) and machine translation (MT) data.

	Avg Sent-Level Document Correlation	Doc-Level Correlation
BLEU	0.5834	0.6312
COMET	0.8343	0.6626
MetricX	0.7635	0.5765

Table 5: For the simultaneous interpretation (SI) data, we derive document-level metric scores for BLEU, COMET, and MetricX in two ways: (1) by computing the average of sentence-level metric scores across the document, and (2) by applying the metrics to the entire document. The human rating for each document is calculated as the average of all its sentence ratings. We then calculate Pearson's correlation between each document-level metric and the human adequacy ratings.

mental designs detailed in Section 3.4. As we saw in Table 4, COMET and MetricX correlate similarly well with human judgments of fluency and adequacy, outperforming all other metrics; when measuring metric sensitivity to the four potential metric failures in Section 4.2.2 and Section 4.2.3, we focus on the MetricX metric for brevity and clarity.

4.2.1 Segmentation

One issue that we observe in the interpretation data is the presence of segmentation errors. Interpreters may break the speech into smaller segments and/or translate them into separate sentences. Although the machine translation system translates each verbatim transcript sentence into a translation sentence, it may still have a different number of sentences than the reference. We find that in the interpretation data, there are 11 documents where the ratio of interpreter sentences to reference sentences is greater than or equal to 1.25, while in the machine translation, there are only 6 documents with a sentence ratio greater than or equal to 1.25. Segmentation differences pose a challenge to the performance of MT metrics, because the metrics often expect a one-to-one alignment between hypothesis and reference sentences. Other datasets face the same issues of segmentation; for example, we observe similar issues in the NAIST (Doi et al., 2021) and VoxPopuli (Wang et al., 2021) datasets.

To see whether metrics are sensitive to these segmentation issues, we employ metrics which are appropriate for both sentence and document-level evaluations: BLEU, COMET, and MetricX. BLEU has no input length restriction, while COMET and MetricX have a 512-token limit. We exclude the documents exceeding this limit, resulting in a set of 59 documents. For COMET, we compute both average sentence-level scores and document-level scores. Following the findings of Deutsch et al. (2023), we use sentence-level and document-level MetricX models to score each document. For human annotations, we average the scores across all sentences within a document.

Table 5 shows the results on metric sensitivity to segmentation. For the correlation of adequacy, we see BLEU improve, while COMET and MetricX both greatly degrade. This indicates that moving from the sentence-level to the document-level does not necessarily resolve the issue of segmentation in metric score, and the effect of shifting from sentence to document-level evaluation differs substantially by metric. However, segmentation differences pose issues beyond the question of sentence boundary, as segmentation is also associated with omission and summarization (discussed in Section 4.2.3).

4.2.2 Disfluency

Now, we assess the impact of the remaining features (disfluency, summarization, and paraphrasing) on metric scores, with a focus on MetricX. These results are summarized in Table 6.

Minor disfluency arises in the interpretation process as the interpreter either misspeaks or is not yet sure what the speaker will say. An example of minor disfluency is shown in Table 7; the MetricX score for the interpretation is 0.5756 and for the

Data	MetricX
MT	0.7351
MT summarized by PaLM	0.6816
MT paraphrased by PaLM	0.7589
SI	0.5928
SI disfluency removed	0.6217

Table 6: Impact on the MetricX scores from perturbations with different interpretation features to the translation data.

Ref: "The alderman for the region has already travelled to Brussels 3 times and has already completed a good proportion of the schedule of works that was outlined in a hearing held before the committee on Petitions in July."
MT: "the regional councilor has already come 3 times here in Brussels and has already implemented a large part of the 'timeline' which was illustrated during a hearing in July before the petitions committee."
SI: "The regional assessor has been 3 times to Brussels and has already done a fair amount of programme put out during a hearing in July **in the peti- Petitions committee**."

Table 7: Example of minor disfluency–indicated in bold–occurring in the simultaneous interpretation (SI), as well as the corresponding machine translation (MT) and reference (Ref) text.

MT is 0.7035.

To measure the impact of disfluencies, we automatically remove them from interpretations (through the process described in Section 3.4). We find that disfluency removal improves MetricX scores by 3%. While this is a very small change, this does indicate that even imperfect disfluency removal leads to an increase in MetricX score, thus demonstrating that MetricX is in fact sensitive to disfluencies.

Again, though only a small change in MetricX score results from the presence of disfluencies, disfluencies can easily be mitigated with disfluency removal, and as they are an organic part of the live interpretation process which do not affect meaning, we argue that these disfluencies should be resolved prior to evaluation. The presence of these disfluencies does not impact the meaning of the interpretation, and we do not expect the machine interpretations to need to produce disfluencies. We also recommend that when creating interpretation datasets, the data curators clean up disfluencies during transcription, or alternatively annotate the disfluencies as in the NAIST dataset (Doi et al., 2021).

4.2.3 Summarization and Paraphrasing

In addition to issues of segmentation and disfluency, we also noted instances of summarization and paraphrasing affecting metric scores.

One such example of summarization can be found in Table 8, for which the interpretation MetricX score is 0.6485 and the MT MetricX score is 0.7710.

Paraphrasing also appears to affect MetricX score, such as in Table 9, where the MetricX score for the interpretation is 0.7171 and for the MT is 0.8215.

To quantify the impact of summarization and paraphrasing on MetricX, we use LLMs to add summarization and paraphrasing to non-simultaneous machine translations as described in Section 3.4, and then observe the impact on MetricX score. The results for this experiment are as shown in Table 6.

Our results indicate that summarization does have a notable impact on MetricX score. Without summarization, the average MetricX score was 0.7351 and after applying summarization this drops to 0.6816. Table 10 breaks the scores down by amount of summarization. We measure summarization via sentence compression ratio, defined as token count in the translation divided by token count in the reference (using the NLTK tokenizer). Interestingly, we find that more summarization leads to a more diminished MetricX score, further confirming that summarization is a weakness of MetricX when evaluating interpretation.

We argue that if no meaning is lost, interpretation metrics should not penalize summarization, as this is again a necessary feature of interpretation, and this therefore needs to be addressed. Still, it is worth noting that we are not able to guarantee that there is no loss of information due to summarization. While our results of sentence compression ratio do indicate the impact of token count on MetricX score, it is possible that in some cases, meaningful information is lost.

When performing the same experiment for paraphrasing, we find that MetricX does handle paraphrasing well, as one would hope. The original MT MetricX score was 0.7351, and after applying paraphrasing via the PaLM model, the MetricX score was 0.7589. Given that paraphrasing actually results in a *higher* MetricX score, paraphrasing is not an issue facing MetricX for interpretation evaluation. Therefore, these sets of experiment indicate that while summarization does pose an issue for MT metrics (in particular with regard to evaluation of interpretation data), paraphrasing does not.

Ref: "The fact that the crisis has hit Naples while the situation is very different in the rest of Italy, for example, in my region, Veneto, where separate collection has been taking place for years without any problems and with a very high recycling rate, means that the **responsibility for the crisis lies with Campanian policy making, with local government officials and, above all, with the serious collusion with the underworld**, which as always sought and made **huge profits from the waste business thanks to Camorra's infiltrating local policy making and local government**."
MT: "If the emergency hit Naples while things are going very differently in the rest of Italy, for example in my region, Veneto, where separate waste collection has been done for years without problems and with a very high recycling rate, it means that the **responsibilities of 'emergency falls on politics and local administrators and, above all, on the heavy connivance with the underworld** which has always sought and obtained **huge profits from the waste business thanks to the infiltration of the Camorra in politics and local administrations**."
SI: "It means that the **responsibility is due to local administration in Campania and operation with criminal elements** that are obtaining **big profits through the in- infiltration of the Camorra into local authorities and government**."

Table 8: Example of summarization–indicated in bold–occurring in the simultaneous interpretation (SI), as well as the corresponding machine translation (MT) and reference (Ref).

4.3 Fine-tuned Metrics for Interpretation Assessment

In order to address our third research question (can we develop a refined automatic metric which achieves even higher correlation with human judgments), we present a pilot experiment that makes use of fine-tuning for interpretation quality assessment. We utilize our z-normalized human annotation scores (from Section 3.2) along with the interpretation and reference pairs to fine-tune a MetricX model. We employ 3-fold cross-validation for our fine-tuning experiments. In each fold, 33% of the annotated data is held out as the test set, while the remaining 67% is used to fine-tune the model. The average correlation across all three folds is reported in Table 11, marked with asterisks. We avoid fine-tuning on MT annotations to ensure the models are directed towards the task of interpretation evaluation. We do additionally apply our fine-tuned models to MT data and report the resulting correlations.

We take two approaches to fine-tuning the base MetricX model: (1) directly fine-tune the base model with our human annotations, and (2) first fine-tune with the DA data from WMT, and then fine-tune with our annotations. We use either adequacy or fluency score to fine-tune the model. The results can be found in Table 11.

Ref: "We set out to achieve the goal of recognising the right of all patients to cross-border healthcare, thus preventing medical tourism."
MT: "The goal which we have tried to achieve is to recognize all patients the right to cross-border healthcare, avoiding healthcare tourism."
SI: "The objective which we were striving towards was to recognise for all patients the right to cross-border healthcare, but avoiding medical tourism."

Table 9: Example of paraphrasing occurring between the simultaneous interpretation (SI) and reference (Ref), plus the corresponding machine translation (MT).

Sentence Compression	MetricX
Overall	0.6816
$Ratio \leq 0.25$	0.5456
$0.25 < Ratio \leq 0.5$	0.5950
$0.5 < Ratio \leq 0.75$	0.6824
$0.75 < Ratio$	0.7419

Table 10: Summarization ULM experiment and MT MetricX after summarization.

For adequacy assessment, we find that the fine-tuned models correlate better with human judgments than off-the-shelf MT metrics. The WMT DA data is helpful in this case. The highest correlation for the interpretation data is achieved by fine-tuning the "DA 15-20 z clipped" model from Juraska et al. (2023) on our z-normalized human annotations. As for fluency, the fine-tuned models also achieve higher correlation with human ratings. However, for fluency, we find that fine-tuning with the DA data does not lead to improved correlation with human judgments. This demonstrates that with just a very small amount of human annotation, we can create a reasonable metric to evaluate interpretation quality. This suggests that future work can make use of quality-annotated interpretation data to overcome the barriers to interpretation data that we have outlined, thus accounting for features commonly found in high-quality interpretations which affect metric scores.

5 Conclusion

In this work, we have performed extensive qualitative and quantitative experimentation to measure the impact of common features of interpretation on metric scores.

We have studied the sensitivity of MT metrics to interpretation features, including disfluency, seg-

Metric	SI Fluency	SI Adequacy	MT Fluency	MT Adequacy
MetricX	0.2988	0.6178	0.1595	0.3133
COMET	0.3011	0.6211	0.1466	0.3422
mT5 + Adequacy ratings		0.6718*		0.4989
mT5 + DA + Adequacy ratings		**0.7031***		0.4528
mT5 + Fluency ratings	**0.4067***		0.2325	
mT5 + DA + Fluency ratings	0.4017*		0.1023	

Table 11: Pearson's correlation between metric scores and human judgments of fluency and adequacy for the simultaneous interpretation (SI) and machine translation (MT) data. The last four rows show the performance of our fine-tuned models. The base model (mT5) is fine-tuned with either adequacy or fluency human ratings, and optionally we fine-tune the base model with DA scores as the first stage fine-tuning. Asterisks indicate the average correlation across all three folds of cross-validation (described in Section 4.3).

mentation, summarization, and paraphrasing. We argue that common interpreter features should not be penalized if the original gist is successfully conveyed, and we find that off-the-shelf MT metrics are indeed sensitive to disfluency and summarization.

Our evaluation shows that the quality of human interpretations is worse than machine translations according to both automatic MT metrics and human evaluation. The low scores are caused not only by the sensitivity of MT metrics to interpretation features (as demonstrated in Section 4.2), but also by persistent errors made by interpreters (as illustrated in Section 4.1). Given this finding, though recent work has argued that human interpretations should be used as references in simultaneous interpretation evaluation (Zhao et al., 2021), we advise against using existing interpretations as references for evaluation. Better data collection procedures and annotations are required to ensure that the interpretation data is of high quality.

Ultimately, though prior work has assumed the functionality of MT metrics for evaluating interpretation data, our findings reveal that minor disfluencies and summarization are unduly punished by existing metrics. In order to perform an accurate evaluation of interpretation data, these features must be addressed.

We propose using fine-tuned learned metrics to assess interpretation quality. With human annotations, even flawed interpretation data can be used to fine-tune a model. As our results show, we are able achieve higher correlation with human judgments using our fine-tuned models than the state-of-the-art MT metrics.

Limitations

While our work provides critical insights into barriers to evaluation of interpretation data and introduces a new metric which accounts for these barriers, it is important to note that our results are on English data. Future work extending our experiments to other languages and domains will give indication into how our insights can be extrapolated to other languages.

Acknowledgements

We thank anonymous reviewers for their feedback.

References

Raja Al-Khanji, Said El-Shiyab, and Riyadh Hussein. 2000. On the use of compensatory strategies in simultaneous interpretation. *Meta*, 45(3):548–557.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan

Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Silvia Bernardini, Adriano Ferraresi, and Maja Milicevic. 2016. From epic to eptic — exploring simplification in interpreting and translation from an intermodal perspective. *Target*, 28:61–86.

Edward Loper Bird, Steven and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Large-scale English-Japanese simultaneous interpretation corpus: Construction and analyses with sentence-aligned data. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 226–235, Bangkok, Thailand (online). Association for Computational Linguistics.

Claudio Fantinuoli and Bianca Prandi. 2021. Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 245–254, Bangkok, Thailand (online). Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chikiu 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 (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 971–976, San Diego, California. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics.

Xiaolei Lu and Chao Han. 2023. Automatic assessment of spoken-language interpreting based on machine-translation evaluation metrics: A multiscenario exploratory study. *Interpreting*, 25(1):109–143.

Dominik Machácek, Matús Zilinec, and Ondrej Bojar. 2021. Lost in interpreting: Speech translation from source or interpreter? In *Interspeech*.

Dominik Macháček, Ondřej Bojar, and Raj Dabre. 2023. MT metrics correlate with human ratings of simultaneous speech translation.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Barbara Moser-Mercer, Alexander Künzli, and Marina Korac. 1998. Prolonged turns in interpreting: Effects on quality, physiological and psychological stress (pilot study). *Interpreting*, 3(1):47–64.

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.

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

Johann C. Rocholl, Victoria Zayats, Daniel David Walker, Noah B. Murad, Aaron Schneider, and Daniel J. Liebling. 2021. Disfluency detection with unlabeled data and small BERT models. In *Interspeech*.

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

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Collection of a simultaneous translation corpus for comparative analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 670–673, Reykjavik, Iceland. European Language Resources Association (ELRA).

Craig Stewart, Nikolai Vogler, Junjie Hu, Jordan Boyd-Graber, and Graham Neubig. 2018. Automatic estimation of simultaneous interpreter performance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 662–666, Melbourne, Australia. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Hao Xiong, Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Dutongchuan: Context-aware translation model for simultaneous interpreting.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021. BSTC: A large-scale Chinese-English speech translation dataset. In *Proceedings of the Second Workshop on Automatic Simultaneous Translation*, pages 28–35, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jinming Zhao, Philip Arthur, Gholamreza Haffari, Trevor Cohn, and Ehsan Shareghi. 2021. It is not as good as you think! evaluating simultaneous machine translation on interpretation data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6707–6715.

Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, Jiahong Yuan, Kenneth Church, and Liang Huang. 2020. Fluent and low-latency simultaneous speech-to-speech translation with self-adaptive training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3928–3937, Online. Association for Computational Linguistics.