ParBLEU: Augmenting Metrics with Automatic Paraphrases for the WMT'20 Metrics Shared Task

Rachel Bawden, Biao Zhang, Andre Tättar, Matt Post

▶ **To cite this version:**

Rachel Bawden, Biao Zhang, Andre Tättar, Matt Post. ParBLEU: Augmenting Metrics with Automatic Paraphrases for the WMT'20 Metrics Shared Task. 5th Conference on Machine Translation, Nov 2020, Online, Unknown Region. ⟨hal-02981143⟩

HAL Id: hal-02981143

https://hal.science/hal-02981143v1

Submitted on 27 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

ParBLEU: Augmenting Metrics with Automatic Paraphrases for the WMT'20 Metrics Shared Task

Rachel Bawden[1] **Biao Zhang**[1] **Andre Tättar**[2] **Matt Post**[3]
[1]School of Informatics, University of Edinburgh, Scotland
[2]University of Tartu, Tartu, Estonia
[3]Johns Hopkins University, Baltimore, Maryland, USA

Abstract

We describe parBLEU, parCHRF++, and parESIM, which augment baseline metrics with automatically generated paraphrases produced by PRISM (Thompson and Post, 2020a), a multilingual neural machine translation system. We build on recent work studying how to improve BLEU by using diverse automatically paraphrased references (Bawden et al., 2020), extending experiments to the multilingual setting for the WMT2020 metrics shared task and for three base metrics. We compare their capacity to exploit up to 100 additional synthetic references. We find that gains are possible when using additional, automatically paraphrased references, although they are not systematic. However, segment-level correlations, particularly into English, are improved for all three metrics and even with higher numbers of paraphrased references.

1 Introduction

One of the major challenges faced when automatically evaluating machine translation (MT) outputs is that there are almost always multiple correct translations of a sentence, and an automatic metric should be able to reward them all. Some of the most widely used MT metrics, including BLEU (Papineni et al., 2002) and CHRF++ (Popović, 2015, 2017), rely on a surface-form comparison of MT outputs to a human-produced reference translation. Both metrics support the use of multiple references. However, even for metrics that support multiple references, human-produced references are expensive to produce and so are rarely available. To overcome this problem, metrics that do not rely on the surface form of reference translations have been developed. One example is ESIM (Chen et al., 2017; Mathur et al., 2019), which uses contextual embeddings with the aim of creating an abstract meaning representation of the reference, with the potential of covering all translations with the correct meaning.

We explore an alternative way of increasing the capacity of MT metrics to reward multiple valid translations: create additional references by automatically paraphrasing the original reference. There have been previous efforts to provide some sort of paraphrase support, mostly concentrating on synonyms (Banerjee and Lavie, 2005; Kauchak and Barzilay, 2006; Denkowski and Lavie, 2014). However, we base our work on a more recent attempt to improve BLEU using diverse automatic parahrasing with high quality MT-style *sentential* paraphrasing (Bawden et al., 2020).

We put this to the test in the WMT'20 metrics shared task by applying Bawden et al.'s (2020) approach to three different metrics: BLEU, CHRF++ and ESIM. We compare the different metrics' capacity to exploit automatically generated multiple references. We choose to use diverse paraphrases produced using PRISM (Thompson and Post, 2020a), since they are available in multiple languages, including most languages of the WMT shared task. We find that gains in correlation are possible, but this depends largely on the language direction and on whether the metric is system- or segment-level. The most positive gains are seen at the segment level, especially for into-English and even at higher numbers of additional paraphrases. This holds for all three metrics, despite ESIM relying on more abstract semantic representations.

2 Overview of Base Metrics

In an extension of (Bawden et al., 2020), we augment three base metrics with automatic paraphrasing. The metrics vary in the basic units of comparison between MT outputs and the reference. BLEU and CHRF++ compare surface representations, BLEU at the token level, whereas CHRF++ also takes into account character n-grams. ESIM is an embedding-based metric, which aims to cap-

ture the semantic relatedness of the sentences. A description of each base metric can be found below.

2.1 BLEU

BLEU (Papineni et al., 2002) is the dominant metric in MT. It is a modified form of n-gram precision, calculated by averaging token n-gram precisions (p_n, $n = 1..4$) and multiplying by a brevity penalty (BP) used to penalise overly short translations:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (1)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \quad (2)$$

$$p_n = \frac{\sum_{h \in H} \sum_{\text{ngram} \in h} \#_{\text{clip}}\left(ngram\right)}{\sum_{h' \in H} \sum_{\text{ngram'} \in h'} \#\left(ngram'\right)}, \quad (3)$$

where c and r are the lengths of the hypothesis and reference sets respectively, H is the set of hypothesis translations, $\#\left(ngram\right)$ the number of times *ngram* appears in the hypothesis, and $\#_{\text{clip}}(ngram)$ is the same but clipped to the maximum number of times it appears in any one reference (if several references are available).

BLEU is typically used in its corpus-based variant, where a single score is produced for a test set. However, a segment-level variant also exists, where each sentence is scored individually. Smoothing is necessary in this segment-level variant to counteract the effect of 0 n-gram precision.

We use the sacreBLEU implementation[1] of BLEU (Post, 2018), with default tokenisation (and `-tok zh` tokenisation for Chinese) and exponential smoothing for the sentence-level variant.

2.2 CHRF++

CHRF++ (Popović, 2017),[2] like BLEU, is a surface-based metric, but which relies on overlap in character n-grams as well as token n-grams (hence its name 'character n-gram F-score'). This theoretically gives it an advantage over BLEU, since it is able to reward partial token matches thanks to its character-level component.

The original chrF (Popović, 2015) was calculated as follows using just character-level n-grams:

$$ngrF\beta = (1 + \beta^2)\frac{ngrP \cdot ngrR}{\beta^2 \cdot ngrP + ngrR},$$

where $ngrP$ and $ngrR$ respectively stand for the arithmetic average of n-gram precision and recall over character n-grams from 1 to N, where β gives more or less weight to precision than recall. CHRF++ expands on this original metric by also including token-level n-grams in this calculation with n from 1 to M. The best results were found with $N = 6$ and $M = 1$ or 2. We use the settings used in the WMT19 shared task: $N = 6$, $M = 1$ and $\beta = 3$. Like BLEU, CHRF++ has a specific corpus-level and sentence-level variant.

2.3 ESIM

ESIM (Chen et al., 2017; Mathur et al., 2019), is an embedding-based metric, which relies on neural models to handle inter-sentence semantic relatedness, going beyond surface-level matching (as in BLEU and CHRF++). ESIM was originally proposed to compare and match sentence pairs for natural language inference (Chen et al., 2017). Mathur et al. (2019) adapted it to evaluate MT performance by pairing the human reference and the MT output as ESIM input. Following (Mathur et al., 2019),[3] we treat the evaluation task as a regression task, and train ESIM models on segment-level human judgments. We train ESIM on the WMT18 metric data for WMT19 evaluation, and WMT18+WMT19 metric data for WMT20 evaluation. ESIM is a sentence-level metric. Scores are averaged to produce a single score for a given corpus.

3 Experiment Setup

Paraphrase generation We use the PRISM system to generate paraphrases. PRISM is a many-many multilingual NMT system covering 39 languages, including all of those of WMT 2019, except Gujarati. In their submission to the WMT 2020 Metrics task, Thompson and Post (2020c) retrained PRISM with five additional languages: Gujarati (for WMT'19), and Inuktitut, Khmer, Pashto, and Tamil (for WMT'20). This provided almost complete coverage of the WMT 2019 and 2020 languages. We use this same model.

By design, PRISM approaches paraphrasing as a zero-shot translation task. As a result, while good for scoring, it is not a particularly good generative model, in terms of being able to produce diverse outputs. Thompson and Post (2020b) have tried to address this, but their implementation does not

[1] https://github.com/mjpost/sacrebleu
[2] https://github.com/m-popovic/chrF/
[3] https://github.com/nitikam/mteval-in-context

We reached a pretty quick agreement, Kouki said.
We reached a fairly quick agreement, Kouki said.
We reached a fairly quick agreement, Kouki was quoted as saying.
We reached a fairly quick agreement, Kouki said
We reached a fairly quick agreement, Kouki was quoted as telling reporters.
We reached a fairly quick agreement, Kouki was quoted to be quoted as saying.
We reached a fairly quick agreement, Kouki was quoted as adding.

Jamsen says the church bells don't ring because of a malfunction.
Jamsen says the church bells do not ring because of a malfunction.
Jamsen says the church bells do not ring because of a maloperation.
Jamsen says the church bells aren't ringing because of an improper functioning.
Jamsen says that the church bells don't ring because of a mal-function.
It's a technical malfunctory, to say it's a technical malfunction.
It's a technical malfunction, I'm sure.

Figure 1: Two examples of automatic paraphrasing from fi–en WMT'20 (original references in bold).

produce n-best lists. We therefore produce n-best lists from the model using Fairseq's built-in diverse beam search tool. For every reference in the WMT19 and WMT20 news test sets, we generate a 100-best list (Vijayakumar et al., 2016).[4]

Figure 1 shows examples of paraphrases of two fi–en WMT'20 references. Note that the paraphrases are divere and generally of high quality. However, the later paraphrases may be noisier.

Integrating multiple references We augment each of the base metrics described in Section 2 to produce three new metrics: parBLEU, parCHRF++ and parESIM. Both BLEU and CHRF++ have inbuilt support for multiple references. For ESIM, we calculate the score for each reference separately and then average them to get the final score.

Metrics Task Setup Awaiting the gold judgments for WMT'20, we test and report the results of each method on the WMT19 metrics task.[5] We follow the metrics task setup (Ma et al., 2019) by calculating the correlation with manual direct assessments (DA) of MT quality (Graham et al., 2013). System-level scores are evaluated using Pearson's r and segment-level correlations using Kendall's τ on the DA assessments converted into relative rankings. Statistically significant improvements (over the single-reference base metric) are marked in bold (with $p \leq 0.05$). Significance is calculated using the Williams test (Williams, 1959) at the system level and bootstrap resampling at the segment level.

[4]We pass the following arguments: `fairseq-interactive ... --beam 100 --nbest 100 --diverse-beam-groups 10 --diverse-beam-strength 1`

[5]http://statmt.org/wmt19/results.html

4 Results

The results are reported in the following three sections for each paraphrase-augmented metric: parBLEU (Section 4.1), parCHRF++ (Section 4.2) and parESIM (Section 4.3). We report results for up to 100 additional paraphrased references, except for parESIM, where we report up to 50 additional references due to the length of time needed to calculate results. There are some general trends:

- There is often a difference between into- and from-English language directions, with more positive results seen into English. This could be due to the potential better quality of the English paraphrases.
- Results are better for into-English at the segment-level, where adding paraphrases tends to help even with more paraphrases.

4.1 parBLEU

Results for parBLEU are found in Table 1 (system-level) and Table 2 (segment-level).

#extra	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
0	0.988	0.959	0.970	0.736	0.849	0.989	0.968	0.901
1	0.986	0.954	0.968	0.737	0.876	0.982	0.977	0.941
2	0.986	0.953	0.968	0.738	0.875	0.981	0.979	0.938
5	0.986	0.954	0.968	0.738	0.879	0.980	0.980	0.933
25	0.984	0.958	0.969	0.739	0.883	0.976	0.982	0.927
50	0.982	0.959	0.969	0.740	0.887	0.974	0.982	0.924
100	0.977	0.957	0.965	0.743	0.888	0.973	0.982	0.897

(a) From-English language directions

#extra	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
0	0.891	0.986	0.798	0.943	0.969	0.861	0.888
1	**0.905**	0.987	0.802	0.951	**0.975**	**0.887**	**0.898**
2	**0.906**	0.987	0.797	0.953	**0.977**	**0.893**	0.894
5	**0.912**	0.987	0.794	**0.955**	**0.981**	**0.897**	0.892
25	**0.925**	0.985	0.784	**0.966**	**0.984**	**0.898**	0.894
50	**0.930**	0.984	0.780	**0.971**	**0.986**	**0.906**	0.892
100	**0.940**	0.979	0.777	**0.977**	**0.990**	**0.919**	0.874

(b) To-English language directions

Table 1: parBLEU system-level results.

System-level results are variable, with a notable difference between into-English and from-English language directions. For a couple of from-English languages, there are some slightly higher correlations but these are not significant, and some deteriorations can be seen when adding paraphrase for others. Adding paraphrased references is more successful for into-English languages. For four of the language directions, adding the maximum number of 100 paraphrases provides the greatest significant correlation gains, suggesting that even more gains could be achieved with more paraphrases. These gains are illustrated in Figure 2a.

#extra	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
0	0.351	0.239	0.381	0.436	0.362	0.309	0.462	0.262
1	**0.359**	**0.259**	**0.409**	0.428	**0.370**	**0.322**	**0.483**	-0.313
2	**0.361**	**0.260**	**0.412**	0.426	0.370	**0.318**	**0.485**	-0.309
5	**0.360**	**0.260**	**0.416**	0.422	0.365	0.311	**0.487**	-0.309
25	**0.366**	**0.265**	**0.422**	0.415	0.362	0.314	**0.489**	0.263
50	**0.362**	**0.267**	**0.425**	0.414	0.357	0.318	**0.489**	0.266
100	**0.369**	**0.268**	**0.423**	0.398	0.333	**0.327**	**0.488**	-0.280

(a) From-English language directions

#extra	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
0	0.050	0.223	0.166	0.363	0.248	0.106	0.312
1	**0.055**	**0.227**	**0.175**	0.367	**0.264**	**0.113**	**0.321**
2	**0.054**	0.226	**0.178**	0.362	**0.268**	**0.114**	**0.320**
5	**0.056**	0.227	**0.181**	0.363	**0.272**	**0.112**	0.317
25	**0.065**	**0.235**	**0.185**	**0.372**	**0.275**	**0.126**	**0.321**
50	**0.070**	**0.241**	**0.186**	**0.375**	**0.284**	**0.127**	**0.324**
100	**0.066**	**0.243**	**0.191**	**0.371**	**0.293**	**0.134**	0.311

(b) To-English language directions

Table 2: parBLEU segment-level results.

Segment-level results show variability according to the language direction too. The greatest gains are seen for the into-English directions, and the highest scores are achieved for the higher order numbers of paraphrases. Some gains are seen for most from-English directions, even with higher numbers of paraphrases. Interestingly, the language directions that see gains at the segment level are not correlated with those that see gains at the system level.

4.2 parCHRF++

System-level and segment-level results can be found in Table 3 and Table 4 respectively. The CHRF++ baseline (0 extra references) is higher than the BLEU baseline for into-English at the system-level and into all languages (except Chinese) at the segment-level.

#extra	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
0	0.984	0.977	0.981	0.836	0.967	0.969	0.985	0.801
1	0.981	0.977	0.979	0.836	0.972	0.967	0.986	0.821
2	0.980	0.977	0.978	0.835	0.972	0.966	0.986	0.824
5	0.980	0.977	0.977	0.835	0.973	0.965	0.986	0.827
10	0.979	0.977	0.977	0.835	0.973	0.965	0.986	0.827
25	0.979	0.976	0.976	0.835	0.974	0.964	0.986	0.823
50	0.979	0.976	0.975	0.835	0.974	0.963	0.985	0.821
100	0.974	0.976	0.972	0.835	0.973	0.962	0.986	0.823

(a) From-English language directions

#extra	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
0	0.909	0.991	0.947	0.966	0.936	0.918	0.955
1	**0.919**	0.991	0.948	0.967	**0.948**	**0.930**	**0.961**
2	**0.922**	0.991	0.948	0.968	**0.950**	**0.932**	**0.960**
5	**0.925**	0.991	0.948	0.969	**0.952**	**0.936**	**0.961**
25	**0.930**	0.991	0.952	**0.972**	**0.952**	**0.940**	**0.962**
50	**0.933**	0.991	0.953	**0.973**	**0.953**	**0.942**	**0.962**
100	**0.938**	0.990	0.950	**0.982**	**0.963**	**0.949**	0.963

(b) To-English language directions

Table 3: parCHRF++ system-level results.

At the system level, as with parBLEU, greater gains are seen for into-English than from-English language directions: all into-English language directions bar fi–en show increases. Moreover, most into-English language directions continue to see improvements with higher numbers of references. This trend can be seen in Figure 2b.

#extra	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
0	0.449	0.323	0.518	0.546	0.497	0.439	0.548	0.238
1	**0.455**	**0.326**	0.519	0.546	0.498	0.431	**0.564**	0.237
5	0.448	0.325	0.517	0.546	0.490	0.418	**0.555**	0.221
25	0.444	**0.327**	0.515	0.545	0.487	0.416	**0.560**	0.209
50	0.443	**0.327**	0.515	0.545	0.485	0.417	**0.559**	0.203
100	0.433	0.322	0.506	0.545	0.459	0.405	0.546	0.193

(a) From-English language directions

#extra	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
0	0.125	0.288	0.254	0.393	0.303	0.182	0.373
1	0.123	0.288	**0.258**	0.396	**0.311**	0.182	0.375
5	0.126	0.286	**0.261**	**0.398**	**0.317**	0.182	0.377
25	**0.129**	0.291	**0.263**	**0.398**	**0.322**	0.183	0.375
50	0.128	0.291	**0.265**	0.397	**0.327**	0.182	0.378
100	0.120	0.285	**0.269**	0.397	**0.313**	0.180	0.367

(b) To-English language directions

Table 4: parCHRF++ segment-level results.

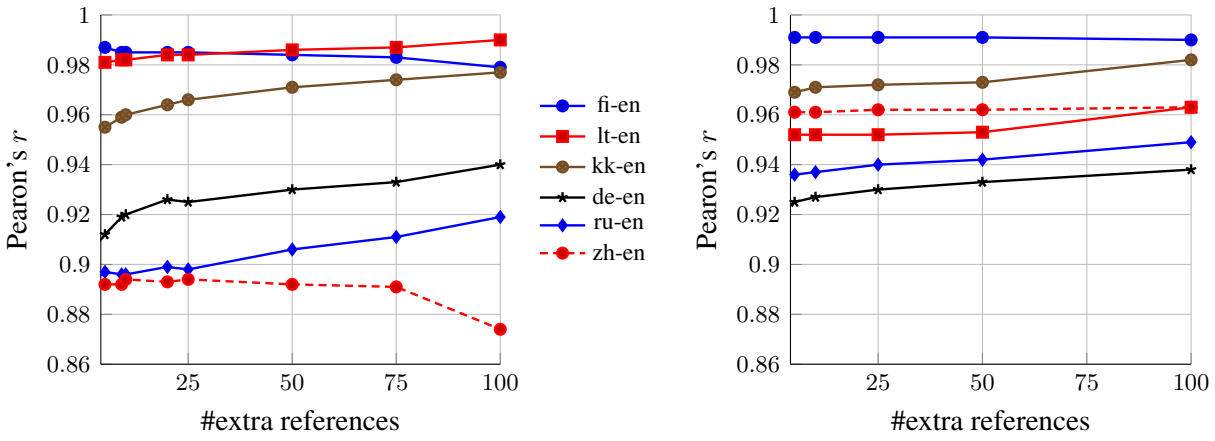
At the segment level, extra references help all into-English directions, although this does depend on the number of references added for some language directions. From-English, some slight gains are seen but in most cases, adding extra references degrades results. At the segment level, the best results can be seen with just one additional reference.

4.3 parESIM

System-level and segment-level results can be found in Table 5 and Table 6 respectively. As an automatic metric that relies on comparing continuous representations (aiming to abstract away from surface forms), we would expect paraphrases to help ESIM less than the two other metrics, for which surface form variation is one of the major limitations.

At the system level, additional paraphrases does not seem to help for any of the language directions, and is even harmful (decreasing correlations as the number of paraphrases is increased). This could be due to the addition of noise in the results, which treats semantically divergent hypotheses as valid. Note however that the correlations start from a strong base—baseline ESIM has a much higher correlation than BLEU and CHRF++.

The segment-level results are more positive: paraphrasing significantly helps four from-English directions (into cs, de, ru and zh). It brings even more positive gains for the into-English language directions, where the best results are often achieved with the higher numbers of additional paraphrases.



(a) parBLEU (b) parCHRF++

Figure 2: System-level results for into-English language directions.

#extra	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
0	0.924	0.990	0.943	0.946	0.978	0.960	0.978	0.942
1	0.918	0.988	0.935	0.942	0.972	0.941	0.977	0.947
2	0.916	0.987	0.932	0.935	0.970	0.932	0.977	0.949
5	0.913	0.986	0.928	0.916	0.967	0.921	0.975	0.949
10	0.912	0.985	0.926	0.895	0.966	0.916	0.975	0.949
25	0.910	0.985	0.924	0.870	0.966	0.912	0.973	0.949
50	0.909	0.984	0.923	0.857	0.966	0.910	0.973	0.950

(a) From-English language directions

#extra	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
0	0.938	0.967	0.875	0.982	0.987	0.973	0.988
1	0.937	0.967	0.873	0.980	0.987	0.972	0.989
2	0.937	0.967	0.873	0.980	0.987	0.971	0.989
5	0.936	0.966	0.872	0.978	0.987	0.971	0.989
10	0.935	0.966	0.872	0.976	0.987	0.971	0.989
25	0.935	0.965	0.871	0.974	0.987	0.970	0.989
50	0.934	0.965	0.870	0.972	0.987	0.969	0.989

(b) To-English language directions

Table 5: parESIM system-level results.

#extra	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
0	0.471	0.356	0.535	0.545	0.508	0.487	0.582	0.330
1	0.475	**0.366**	0.532	0.517	0.494	0.489	**0.593**	**0.352**
2	**0.476**	**0.364**	0.526	0.488	0.476	0.477	**0.593**	**0.348**
5	**0.480**	**0.368**	0.520	0.413	0.452	0.467	**0.596**	**0.344**
10	**0.477**	**0.370**	0.519	0.359	0.436	0.467	**0.595**	**0.345**
25	0.475	**0.370**	0.514	0.307	0.426	0.463	**0.589**	**0.339**
50	0.476	**0.370**	0.515	0.290	0.424	0.464	**0.592**	**0.343**

(a) From-English language directions

#extra	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
0	0.155	0.328	0.294	0.426	0.348	0.190	0.347
1	**0.163**	0.330	0.293	0.424	0.352	0.193	**0.352**
2	**0.168**	0.329	0.293	0.427	**0.353**	0.192	**0.354**
5	**0.174**	0.328	0.294	0.421	**0.354**	**0.199**	**0.355**
10	**0.174**	**0.334**	0.294	0.419	**0.356**	**0.199**	**0.354**
25	**0.175**	**0.334**	0.295	0.415	**0.359**	**0.201**	**0.354**
50	**0.176**	**0.333**	0.294	0.412	**0.357**	**0.200**	**0.356**

(b) To-English language directions

Table 6: parESIM segment-level results.

4.4 Additional parBLEU comparisons

Following the shared task, we explored some alternative versions of parBLEU.

Replacing the original reference Concurrently to Bawden et al. (2020), Freitag et al. (2020) also review paraphrasing for BLEU, although they focus on human paraphrasing. They find that better correlations are achieved by replacing the original reference with a human paraphrased one, as original references often display translationese. We test this observation here, but using our automatic paraphrases. Results are shown in Table 7 (system level) and Table 8 (segment level).

#extra	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
Original	0.988	0.959	0.970	0.736	0.849	0.989	0.968	0.901
Paraphrased	0.978	0.946	0.951	0.115	**0.941**	0.946	0.983	0.936

(a) From-English language directions

Metric	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
Original	0.891	0.986	0.798	0.943	0.969	0.861	0.888
Paraphrased	**0.916**	0.988	0.799	0.952	**0.978**	**0.905**	0.902

(b) To-English language directions

Table 7: parBLEU system-level results when using the original reference versus the first paraphrase.

#extra	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
Original	0.351	0.239	0.381	0.436	0.362	0.309	0.462	0.262
Paraphrased	0.327	0.228	0.342	-0.149	0.224	0.181	0.455	0.210

(a) From-English language directions

Metric	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
Original	0.050	0.223	0.166	0.363	0.248	0.106	0.312
Paraphrased	0.046	0.222	0.167	0.360	0.253	0.106	0.316

(b) To-English language directions

Table 8: parBLEU segment-level results when using the original reference versus the first paraphrase.

We find that replacing the original reference with its first paraphrase results in higher correlations for the into-English language directions at the system

level (although the gain is only significant for three directions), and there do not seem to be gains at the segment level. In general, it harms both correlation types for the from-English language directions, probably due to the better quality of the English paraphrasing compared to that of the other languages. This appears to confirm Freitag et al.'s observation, as long as the quality of the paraphraser is good enough, which is our hypothesis concerning the into-English language directions.

Type of paraphraser We compare three different paraphrasers for the into-English language directions: (i) the 'sampled' diverse paraphrasing approach from (Bawden et al., 2020), (ii) the n-best PRISM paraphrases, and (iii) the n-best diverse PRISM paraphrases used elsewhere in this paper. The results are given in Table 9. Somewhat surprisingly, even though they are not designed to be diverse, the n-best paraphrases give good correlations, at least up to 20 paraphrases, which was the maximum number tested with the sampled paraphraser. The sampled paraphrases also often perform better than the diverse approach produced by the PRISM paraphraser. One reason for this could be that the sampled paraphraser is trained specifically as an English paraphraser, whereas the PRISM paraphraser is multilingual (therefore providing greater support for automatic evaluation).

Exclusion of outliers Mathur et al. (2020) suggested that system-level correlations computed with Pearson's are artificially inflated due to the presence of outliers, which are typically very poorly performing systems with low human scores. They propose a method based on *mean average deviation* (MAD) to exclude those outliers. We applied this method to the WMT19 system-level data to exclude systems, and then recomputed the system-level correlations.

The complete results are in Table 10. Comparing this to Table 7, we see an absolute drop in values, but little to nothing in the way of reversals between the BLEU (single-reference, zero-paraphrase) baseline and the paraphrase methods.

5 Conclusions and Future Work

The goal with any metric is to balance accuracy with ease-of-use. For our submission to the WMT20 metrics task, we extended our work investigating paraphrased English references (Bawden et al., 2020), by using a multilingual paraphraser.

Type	#extra	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
	0	0.891	0.986	0.798	0.943	0.969	0.861	0.888
Sampled	1	**0.915**	0.985	0.801	0.960	**0.984**	**0.907**	0.891
	2	**0.926**	0.986	0.799	0.962	**0.987**	**0.918**	0.896
	10	**0.942**	0.980	0.800	0.970	**0.992**	**0.932**	0.906
	20	**0.946**	0.976	0.800	0.973	**0.992**	0.933	0.907
n-best	1	**0.910**	0.987	0.801	0.952	**0.975**	**0.884**	**0.899**
	2	**0.913**	0.988	0.802	0.954	**0.975**	**0.894**	**0.901**
	10	**0.935**	0.989	0.801	0.959	**0.978**	**0.915**	**0.907**
	20	**0.938**	0.989	0.800	0.960	**0.981**	**0.923**	**0.913**
diverse	1	**0.905**	0.987	0.802	0.951	**0.975**	**0.887**	**0.898**
	2	**0.906**	0.987	0.797	0.953	**0.977**	**0.893**	0.894
	10	**0.061**	0.225	**0.182**	0.369	**0.272**	**0.121**	**0.320**
	20	**0.926**	0.985	0.784	**0.964**	**0.984**	**0.899**	0.893

(a) System-level

Type	#extra	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
	0	0.050	0.223	0.166	0.363	0.248	0.106	0.312
Sampled	1	**0.054**	**0.237**	**0.181**	0.364	**0.282**	**0.121**	0.309
	2	**0.057**	**0.239**	**0.185**	0.367	**0.283**	**0.119**	0.307
	10	**0.078**	**0.254**	**0.191**	**0.376**	**0.302**	**0.127**	0.314
	20	**0.077**	**0.252**	**0.192**	**0.378**	**0.308**	**0.125**	0.316
n-best	1	**0.056**	**0.227**	**0.175**	0.367	**0.261**	**0.111**	**0.324**
	2	**0.054**	0.226	**0.180**	**0.370**	**0.272**	**0.119**	**0.323**
	10	**0.064**	**0.232**	**0.190**	**0.381**	**0.278**	**0.133**	**0.324**
	20	**0.062**	**0.240**	**0.193**	**0.384**	**0.289**	**0.131**	**0.332**
diverse	1	**0.055**	**0.227**	**0.175**	0.367	**0.264**	**0.113**	**0.321**
	2	**0.054**	0.226	**0.178**	0.362	**0.268**	**0.114**	**0.320**
	10	**0.061**	0.225	**0.182**	0.369	**0.272**	**0.121**	**0.320**
	20	**0.063**	**0.234**	**0.183**	**0.371**	**0.273**	**0.124**	**0.323**

(b) Segment-level

Table 9: Correlation results for parBLEU for into-English language directions.

#extra	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
0	0.988	0.828	0.961	0.736	0.591	0.989	0.946	0.901
1	0.986	0.827	0.953	0.737	0.560	0.982	0.964	0.941
2	0.986	0.824	0.953	0.738	0.559	0.981	0.969	0.938
5	0.986	0.826	0.951	0.738	0.563	0.980	0.972	0.933
25	0.984	0.837	0.951	0.739	0.559	0.976	0.972	0.927
50	0.982	0.837	0.948	0.740	0.563	0.974	0.972	0.924
100	0.977	0.821	0.939	0.743	0.530	0.973	0.970	0.897

(a) From-English language directions

#extra	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
0	0.828	0.986	0.967	0.917	0.968	0.844	0.823
1	0.844	0.987	0.970	0.911	**0.975**	**0.876**	0.837
2	0.843	0.987	0.971	0.913	**0.978**	**0.881**	0.827
5	0.851	0.987	0.969	0.910	**0.981**	**0.886**	0.817
25	**0.872**	0.985	0.968	0.925	**0.985**	**0.889**	0.824
50	**0.879**	0.984	0.969	0.924	**0.988**	**0.898**	0.817
100	**0.896**	0.979	0.971	0.885	**0.992**	**0.910**	0.804

(b) To-English language directions

Table 10: System-level results with parBLEU with outlier systems excluded. The *removed* row denotes how many systems were considered to be outliers

One component of ease-of-use, particularly for a metric, is to avoid highly language-specific parameter searches. Our work here used a single model and diversity parameter setting. It is possible that other approaches would yield more success: for example, varying the number of references based on reference length or complexity, or looking at other diverse generation techniques. However they are not guaranteed to work and raise questions about

the usefulness of extending surface-based metrics in the neural age. BLEU is appealing because of its simplicity and universality, but the emerging evidence (cf. Mathur et al. (2020)) suggest that the most promising approach for future work in MT evaluation is in model-based deep-learning approaches. What is encouraging and also somewhat surprising is that the embedding-based ESIM also seems to benefit from the addition of automatically paraphrased references at the segment level, especially into English.

Acknowledgements

This work was supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 825299 (GoURMET), 825303 and the UK Engineering and Physical Sciences Research Council (EPSRC) fellowship grant EP/S001271/1 (MT-Stretch).

References

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.

Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar, and Matt Post. 2020. A Study in Improving BLEU Reference Coverage with Diverse Automatic Paraphrasing. In *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be Guilty but References are not Innocent. arXiv:2004.06063.

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.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, USA. Association for Computational Linguistics.

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

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. 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.

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

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

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.

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

Brian Thompson and Matt Post. 2020a. 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)*, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation (Volume 1: Research Papers)*, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020c. PRISM: The JHU Submission to the Metrics Shared Task at WMT'20. In *Proceedings of the Fifth Conference on Machine Translation (Volume 2: Shared Task Papers)*, Online. Association for Computational Linguistics.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

Evan James Williams. 1959. *Regression Analysis*. Wiley, New York.