Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing

Brian Thompson
Johns Hopkins University
brian.thompson@jhu.edu

Matt Post
Johns Hopkins University
post@cs.jhu.edu

Abstract

We propose the use of a sequence-to-sequence paraphraser for automatic machine translation evaluation. The paraphraser takes a human reference as input and then force-decodes and scores an MT system output. We propose training the aforementioned paraphraser as a multilingual NMT system, treating paraphrasing as a zero-shot "language pair" (e.g., Russian to Russian). We denote our paraphraser "unbiased" because the mode of our model's output probability is centered around a copy of the input sequence, which in our case represent the best case scenario where the MT system output matches a human reference. Our method is simple and intuitive, and our single model (trained in 39 languages) outperforms or statistically ties with all prior metrics on the WMT19 segment-level shared metrics task in all languages, excluding Gujarati where the model had no training data. We also explore using our model conditioned on the source instead of the reference, and find that it outperforms every quality estimation as a metric system from the WMT19 shared task on quality estimation by a statistically significant margin in every language pair.

1 Introduction

Machine Translation (MT) systems have improved dramatically in the past several years. This is largely due to advances in neural MT (NMT) methods (Sutskever et al., 2014; Bahdanau et al., 2015), but the pace of improvement would not have been possible without automatic MT metrics, which provide immediate feedback on MT quality without the time and expense associated with obtaining human judgments of MT output.

However, the improvements that existing automatic metrics helped enable are now causing the correlation between human judgments and automatic metrics to break down (Ma et al., 2019), especially for BLEU (Papineni et al., 2002), which has been the de facto standard metric since its introduction almost two decades ago. The problem currently appears limited to very strong systems, but as hardware, methods, and available training data improve, it is likely BLEU will fail more frequently in the future. This could prove extremely detrimental if the MT community fails to adopt an improved metric, as good ideas could quietly be discarded or rejected from publication because they do not correlate with BLEU. In fact, it is possible this is already happening.

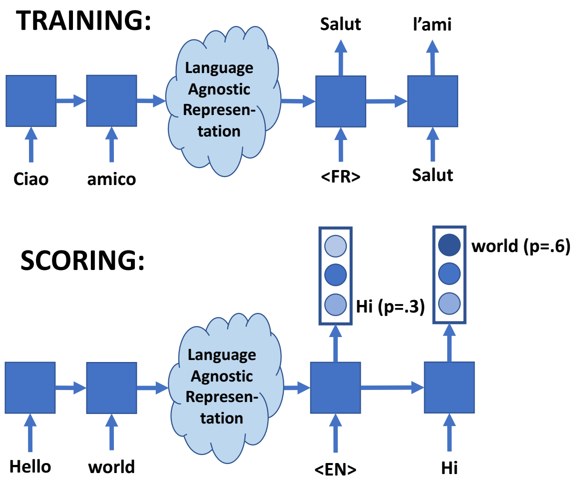

Figure 1: Our model is trained on multilingual parallel examples such as "Ciao amico" translated to French is "Salut l'ami." At evaluation time, the model is used in zero-shot mode to score MT system outputs conditioned on their corresponding human references. For example, the MT system output "Hi world" conditioned on the human reference "Hello world" is found to have token probabilities [0.3, 0.6].

We propose using a sentential, sequence-to-sequence paraphraser to force decode and score MT outputs conditioned on their corresponding human references. Our model effectively stores the entire (exponentially large) set of potential paraphrases of a sentence, both valid and invalid, and we "query" the model with the system output to see how well the system output paraphrases the human reference translation.

The best possible MT output is one which perfectly matches a human reference; therefore, in our application, an ideal paraphraser would be one with an output distribution centered around a copy of its input sentence. We denote such a model an "unbiased paraphraser" to distinguish it from a standard paraphraser trained to produce output which conveys the meaning of the input while also being *lexically and/or syntactically different from it*. For this reason, we propose using a multilingual NMT system as an unbiased paraphraser by treating paraphrasing as a zero-shot "language pair" (e.g., Russian to Russian). We show that a multilingual NMT model is much closer to an ideal unbiased paraphraser than a generative paraphraser trained on synthetic paraphrases. It also allows us to train a single model for all the languages we wish to evaluate.

Figure 1 illustrates our method, which we denote *Prism* (Probability is the metric). Figure 2 shows how our model (see § 4) penalizes both fluency and adequacy errors given a human reference.

We train a single model in 39 languages and show that it outperforms or statistically ties with every metric and baseline from the WMT 2019 MT metrics task (Ma et al., 2019), as well as the recently published BERTscore method (Zhang et al., 2020), at segment-level human correlation in all languages except Gujarati (which was not included in our training data). We find that our model performs well at judging strong NMT systems, as evidenced by positive human correlation on the top four systems (as judged by humans) submitted to WMT19 in every language pair. In contrast, BLEU has negative correlation in 5 language pairs. Additionally, we show that our method can be applied to the task of "Quality estimation (QE) as a metric" (by conditioning on the source instead of the reference) and outperforms all prior methods submitted to the WMT 2019 QE shared task (Fonseca et al., 2019) by a statistically significant margin in every language pair. We release code and models.[1]

Finally, we present analysis which shows that: (1) Due to the effort of the human translators, our multilingual NMT system (which we use as an unbiased paraphraser) need not be SOTA at translation in order to judge SOTA MT systems; (2) Our method has high correlation with human judgments even when those human judgments were made without using the reference; (3) Our unbiased paraphraser outperforms a standard generative paraphraser trained on synthetic paraphrases; and (4) Our method outperforms a sentence embedding-based contrastive semantic similarity approach, which is also trained on bitext in many languages, even when that method is augmented with language model (LM) scores to address fluency.

[1] https://github.com/thompsonb/prism

2 Related Work

MT Metrics Early MT metrics like BLEU (Papineni et al., 2002) and NIST (Doddington, 2002a) use token-level n-gram overlap between the MT output and the human reference. Overlap can also be measured at the character level (Popović, 2015, 2017) or using edit distance (Snover et al., 2006). Many MT metrics consider the semantic similarity of references and translations via word- and/or sentence-level embeddings, including ReVal (Gupta et al., 2015), RUSE (Shimanaka et al., 2018), WMDO (Chow et al., 2019), and ESIM (Mathur et al., 2019). MEANT (Lo and Wu, 2011) and MEANT 2.0 (Lo, 2017) measure similarity between semantic frames and roll fillers. Current SOTA methods including YiSi (Lo, 2019) and BERTscore (Zhang et al., 2019, 2020) rely on contextualized embeddings (Devlin et al., 2019), trained on large (non-parallel) corpora. In contrast, our work exploits parallel bitext.

Paraphrase Databases Prior work has explored exploiting parallel bitext to identify phrase level paraphrases (Bannard and Callison-Burch, 2005; Ganitkevitch et al., 2013) including multilingual bitext (Ganitkevitch and Callison-Burch, 2014). Paraphrase tables have, in turn, been used in MT metrics to reward systems for producing words (Banerjee and Lavie, 2005) or phrases (Zhou et al., 2006; Denkowski and Lavie, 2010) which paraphrase part of the human reference. Our work can be viewed as extending this idea to the sentence level, without having to enumerate the millions or billions of paraphrases (Dreyer and Marcu, 2012) for each sentence.

Multilingual NMT Multilingual NMT (Dong et al., 2015) has been shown to rival performance of single language pair models (Aharoni et al., 2019; Arivazhagan et al., 2019) in high-resource

	Word-level paraphraser log probabilities	H(out|in)	sBLEU	LASER
Copy	Jason went to school at the University of Madrid . <EOS> -0.08 -0.26 -0.16 -0.16 -0.12 -0.11 -0.14 -0.10 -0.10 -0.11 -0.10	-0.13	100.0	1.000
Disfluent	Jason went **school** at **University** of Madrid . <EOS> -0.08 -0.26 **-7.21** -0.12 **-4.81** -0.10 -0.11 -0.11 -0.10	-1.43	35.5	0.989
Inadequate	Jason **will** **go** to school at the University of Madrid . <EOS> -0.08 **-9.77** **-0.76** -0.22 -0.19 -0.14 -0.15 -0.16 -0.10 -0.10 -0.12 -0.10	-0.99	70.8	0.960
	Jason went to school at the University of **Berlin** . <EOS> -0.08 -0.26 -0.16 -0.16 -0.12 -0.11 -0.14 -0.10 **-10.34** -0.12 -0.10	-1.06	78.3	0.957
Fluent & Adequate	Jason **attended** the University of Madrid . <EOS> -0.08 **-2.01** -1.63 -0.42 -0.10 -0.09 -0.16 -0.10	-0.57	41.1	0.918

Figure 2: Example token-level log probabilities from our model for various output sentences, conditioned on input sentence (i.e., human reference) *"Jason went to school at the University of Madrid."*. H(out|in) denotes the average token-level probability. We observe that our model generally penalizes any deviations (**bolded**) from the input sentence, but tends to penalize deviations which change the meaning of the sentence or introduces a disfluency more harshly than those which are fluent and adequate. Sentence-level BLEU with smoothing=1 ("sBLEU") and LASER embedding cosine similarity ("LASER") are shown for comparison. We note that LASER appears fairly insensitive to disfluencies.

languages while also improving low-resource translation via transfer learning from higher-resource languages (Zoph et al., 2016; Nguyen and Chiang, 2017; Neubig and Hu, 2018). An extreme low-resource setting is where the system translates between languages seen during training, but in a language *pair* where it did not see *any* training data, denoted zero-shot translation. Despite some evidence that intermediate representations are not truly language agnostic (Kudugunta et al., 2019), zero-shot translation has been shown successful, especially between related languages (Johnson et al., 2017; Gu et al., 2018; Pham et al., 2019). Our work treats paraphrasing as zero-shot "translation" between (very!) related languages (e.g., Russian to Russian).

Semantic Similarity Parallel corpora in many language pairs has been used to produce fixed-size, multilingual sentence representations (Schwenk and Douze, 2017; Wieting et al., 2017; Artetxe and Schwenk, 2018; Wieting et al., 2019; Raganato et al., 2019). LASER (Artetxe and Schwenk, 2018), for example, trains a variant of NMT with a fixed-size intermediate representation in 93 languages. The decoder is then discarded, and the encoder is used to produce multilingual sentence embeddings. Embeddings can be compared (e.g., with cosine distance) to measure intra- or cross-lingual semantic similarity. We show that our method outperforms LASER embedding similarity, even when LASER is augmented with an LM.

Generative Paraphrasing Sentential paraphrasing can be accomplished by training an MT system on paraphrase examples instead of translation pairs (Quirk et al., 2004). While natural paraphrase datasets do exist (Quirk et al., 2004; Coster and Kauchak, 2011; Fader et al., 2013; Lin et al., 2014), they are somewhat limited.[2] An alternative is to start with much more plentiful bitext and back-translate one side in to the language of the other to create synthetic paraphrases on which to train (Prakash et al., 2016; Wieting and Gimpel, 2018; Hu et al., 2019a,b,c). Tiedemann and Scherrer (2019) propose using paraphrasing as a way to measure the semantic abstraction of multilingual NMT. They also propose using a multilingual NMT model as a generative paraphraser.[3]

3 Method

We propose using a paraphraser to force decode and estimate probabilities of MT system outputs, conditioned on their corresponding human references. Let $p(y_t|y_{i<t}, x)$ be the probability our paraphraser assigns to the t-th token in output sequence y, given the previous output tokens $y_{i<t}$ and the input sequence x. We consider two ways of combining token-level probabilities from the model – sequence-level log probability (G) and average

[2]See Federmann et al. (2019) for analysis comparing methods of generating paraphrases.

[3]We find that generating from a well trained multilingual NMT system tends to produce copies of the input, as opposed to interesting paraphrases – see Appendix A.

token-level log probability (H):

$$G(y|x) = \sum_{t=1}^{|y|} \log p(y_t|y_{i<t}, x)$$
$$H(y|x) = \frac{1}{|y|} G(y|x)$$

Let sys denote MT system output, ref denote human reference, and src denote source. While we expect scoring sys conditioned on ref to be most indicative of the quality of sys, we also explore scoring ref conditioned on sys. This is done because we find qualitatively that output sentences which drop some of the meaning conveyed by the input sentence are penalized less harshly by the model than output sentences which contain extra information not present in the input sentence.[4,5]

We postulate that the output sentence that best represents the meaning of an input sentence is, in fact, simply a copy of the input sentence, as precise word order and choice often convey subtle connotations. As such, we seek a model which we denote an "unbiased paraphraser" whose output distribution is *centered around a copy of the input sentence*. While a standard generative paraphraser is trained to retain semantic meaning, it does not meet our criteria because it is *simultaneously* trained to produce output which is lexically/syntactically different than its input, a key element in generative paraphrasing (Bhagat and Hovy, 2013).

We propose using a multilingual NMT system as an unbiased paraphraser. A multilingual NMT system consists of an encoder which maps a sentence in to an (ideally) language agnostic semantic representation, and decoder to map that representation back to a sentence. The model has only seen bitext in training, but we propose to treat paraphrasing as a zero-shot "language pair" (e.g., Russian to Russian).

As our model is multilingual, we can also score MT system output conditioned on the source instead of the human reference. This task is denoted "quality estimation (QE) as a metric" and was part of the WMT19 QE shared task (Fonseca et al., 2019). We use "Prism-ref" to denote our metric, and "Prism-src" to denote our QE as a metric.

Our final system-level metric and QE metric are defined based on results on our development set (see § 5.2) as follows:

$$\text{Prism-ref} = \frac{1}{2} H(\mathrm{sys}|\mathrm{ref}) + \frac{1}{2} H(\mathrm{ref}|\mathrm{sys})$$
$$\text{Prism-src} = H(\mathrm{sys}|\mathrm{src})$$

To obtain system-level scores, we average segment-level scores over all segments in the test set.

4 Experiments

We train a multilingual NMT model and explore the extent to which it functions as an unbiased paraphraser. We then conduct several preliminary experiments on the WMT18 MT metrics data to determine how to best utilize the token-level probabilities from the paraphraser, and report results on the WMT19 system- and segment-level metrics and QE as metrics tasks.

4.1 Data Preparation

Our method relies heavily on a model, which in turn relies heavily on the data on which it is trained, so we describe here the rationale behind the design decisions made regarding the training data. Full details sufficient for replication are provided in Appendix B. Training a single large model consumed the majority of our compute budget, thus performing ablations, especially on full sizes models, is unfortunately beyond of the scope of this work.

Language Agnostic Representations To encourage our intermediate representation to be as language agnostic as possible, we choose datasets with as much language pair diversity as possible (i.e., not just en-* and *-en), as Kudugunta et al. (2019) has shown that encoder representation is affected by both the source language and target language. While it is common to append the target language token to the source sentence, we instead prepend it to the target sentence so that the encoder cannot do anything target-language specific with this tag. At test time, we force-decode the desired language tag prior to scoring.

Noise NMT systems are known to be sensitive to noise, including sentence alignment errors (Khayrallah and Koehn, 2018), so we perform filtering with LASER (Schwenk, 2018; Chaudhary et al., 2019). We also perform language ID filtering

[4]Scoring in both directions to penalize the presence of information in one sentence but not the other is similar, at least in spirit, to methods which use bi-directional textual entailment as an MT metric (Padó et al., 2009; Khobragade et al., 2019).

[5]Conditional probabilities estimated by MT systems have been shown to be effective at filtering out noisy MT training data (Junczys-Dowmunt, 2018).

using FastText (Joulin et al., 2016) to avoid training the decoder with incorrect language tags.

Number of Languages Aharoni et al. (2019) found that performance of zero-shot translation in a related language pair increased with the number of languages substantially between 5 languages and 25 languages, with a plateau somewhere between 25 and 50 languages. We view paraphrasing as zero-shot translation between sentences in the same language, so we expect to need a similar number of languages.[6]

Copies We filter sentence pairs with excessive copies and partial copies, as multiple studies (Ott et al., 2018; Khayrallah and Koehn, 2018) have noted that MT performance degrades substantially when systems are exposed to copies in training.

4.2 Model Training Details

Our data comes primarily from Wikimatrix (Schwenk et al., 2019), Global Voices,[7] EuroParl (Koehn, 2005), SETimes,[8] and United Nations (Eisele and Chen, 2010). The data processing described above and in Appendix B results in 99.8M sentence pairs in 39 languages.[9] The most common language is English, at 16.7% of our data, while the least common 20 languages account for 21.9%. The full list of languages and amounts is given in Appendix B.

We train a sentencepiece (Kudo and Richardson, 2018) model with a 64k vocabulary size on the concatenation of all data, and filter sentences with length greater than 200 subwords. Multilingual NMT performance increases significantly with the size of the model (Huang et al., 2019), so we train a model as large a feasible given our compute budget constraints. We train a Transformer (Vaswani et al., 2017) in fairseq (Ott et al., 2019) with eight encoder layers, eight decoder layers, an embedding size of 1280, feed forward layer size of 12288, 20 attention heads, learning rate of 0.0004, batch size of 1800 tokens with gradient accumulation over 200 batches, gradient clipping of 1.2, and dropout of 0.1. We train for 6 epochs, which takes approximately 9 days on a `p3.16xlarge` instance rented from Amazon AWS, which has 8 Volta P100 GPUs with 16 GB of memory each. The model has approximately 745M parameters.

4.3 Baselines and Contrastive Methods

We compare to all baselines and submissions to the WMT19 shared metrics task (Ma et al., 2019), as well as BERTscore F1 (Zhang et al., 2019, 2020), which did not submit to or report on WMT19. We explore several contrastive methods to better understand the performance of our method.

Generative Sentential Paraphraser We compare scoring with our unbiased paraphraser vs a standard, English-only paraphraser trained on the Parabank2 dataset (Hu et al., 2019c). We train a Transformer with an 8-layer encoder, 8-layer decoder, 1024 dimensional embeddings, feed-forward size of 8192, and 16 attention heads.

LASER We explore using the cosine distance between LASER embeddings of the MT output and human reference, using the pretrained 93-language model provided by the authors.[10] We are particularly interested in LASER as it, like our model, is trained on parallel bitext in many languages.

Language Model We find qualitatively that LASER is fairly insensitive to disfluencies (see Figure 2), so we also explore augmenting it with language model scores of the system outputs. We train a multilingual language model on the same data as our multilingual NMT system. The model architecture is based on GPT-2 (Radford et al., 2019), and we use the fairseq `transformer_lm_gpt2_small` implementation. We train for 200k updates of approximately 131k tokens. The model has 369M parameters. We train with shared embeddings and a learning rate of 0.0005, and we stop gradients at sentence boundaries, using `--sample-break-mode eos`.

4.4 Paraphraser Bias

We explore the extent to which our paraphraser is unbiased in several ways: Qualitatively, we generate from the model using beam search and examine the output. Quantitatively, we contrast the conditional probabilities of three outputs for the same input: (1) the sequence generated by the model

[6] In preliminary experiments using smaller ("Transformer Big") models, we actually saw similar performance for models trained on 10 and 39 languages. This is perhaps due to our choice of datasets with many language pairs (Aharoni et al. (2019) train on language pairs in and out of English), but we hesitate to draw conclusions from those models as they had significantly worse performance than our full size model.

[7] `casmacat.eu/corpus/global-voices.html`

[8] `nlp.ffzg.hr/resources/corpora/setimes/`

[9] For every sentence pair (a,b) in our 99.8M examples, we train on both (a,b) and (b,a)

[10] `github.com/facebookresearch/LASER`

	Parabank2	This Work
$H(BS\|r0)$	-0.501	-0.225
$H(r0\|r0)$	-1.157	-0.303
$H(r1\|r0)$	-2.246	-2.187
$BLEU(BS, r0)$	31.9	82.8

Table 1: Average token log probability (H) for a sequence generated via beam search (BS), a copy of the input ($r0$), and a high-quality human paraphrase of the input ($r1$), for a generative paraphraser vs our model, conditioned on $r0$ in all cases. BLEU is also computed for the beam search output of each model, with respect to $r0$. Note that BLEU for $r1$ with respect to $r0$ is 17.1.

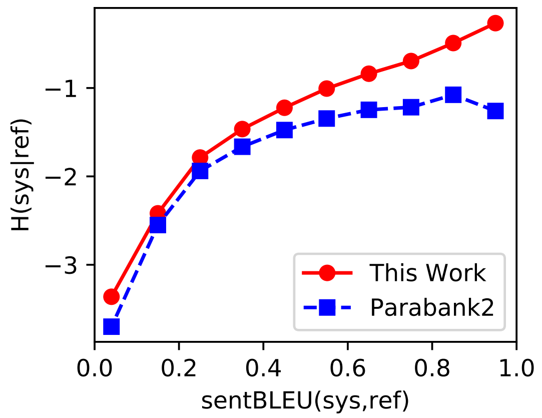

Figure 3: Average model score as a function of average lexical difference, as measured by sentBLEU, for every English (sys,ref) pair submitted to WMT19. Sentence pairs are split in to 10 uniform sentBLEU bins.

via beam search; (2) a copy of the input; and (3) a human paraphrase of the input. We use the English side of the zh-en newstest17 (Bojar et al., 2017) as input, as we can use the second human reference released by Hassan et al. (2018) as a human paraphrase. We also consider model score as a function of lexical similarity (as measured by sentenceBLEU, with smoothing=1) by considering all (system, reference) pairs for all systems submitted to WMT19 in all language pairs translating in to English. We bin the sentence pairs by sentenceBLEU and then consider the average model score for both our model and Parabank2.

4.5 MT Metrics Evaluation

We report segment-level performance with the Kendall's τ variant used in the shared task. System-level performance is computed following the shared task as the Pearson correlation with the mean of the human judgments from Bojar et al. (2018). We employ bootstrap resampling (Koehn, 2004; Graham et al., 2014) following Ma et al. (2018, 2019), using the scripts released for the shared task, to estimate confidence intervals for each metric. Metrics with non-overlapping 95% confidence intervals are identified as having a statistically significant difference in performance.

5 Results

5.1 Paraphraser Bias Results

We find that for our model, a copy of the input is almost as probable as the beam search output (see Table 1). In contrast, the gap is much larger for Parabank2. Additionally, beam search from our model produces output which is very similar to the input (BLEU of 82.8 with respect to input), as desired, while the Parabank2 model tends to change the output more (BLEU of 31.9 with respect to input). This is supported by qualitative inspection, where we see our model tends to produce copies or near copies of the input while the Parabank2 model has a clear tendency to make more significant changes, which occasionally also significantly alters the meaning of the sentence (see Appendix A). We expect that sentBLEU, when averaged over many sentences, should track with semantic similarity, thus our method should track (on average) with sentBLEU as well. We find this to be the case with our multilingual paraphraser, but Parabank2 has nearly the same scores for output with sentBLEU between 0.6 and 1.0 (see Figure 3). All of these findings support our hypothesis that our model is closer to an ideal unbiased paraphraser than the contrastive Parabank2 model which is trained on synthetic paraphrases.

5.2 Preliminary (Development) Results

Preliminary experiments on the WMT18 metrics task data are shown in Figure 4 and Appendix C. We find that length-normalized log probability (H) slightly outperforms un-normalized log probability (G). When using the reference, we find an equal weighting of $H(\text{sys}|\text{ref})$ and $H(\text{ref}|\text{sys})$ to be approximately optimal, but we find that when using the source, $H(\text{src}|\text{sys})$ does not appear to add any useful information to $H(\text{sys}|\text{src})$. These findings were used to select the definitions of Prism-ref and Prism-src in § 3.

We find the probability of sys as estimated by an LM [$H(\text{src})$] and the cosine distance

	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	de-cs	de-fr	fr-de
BERTSCORE (Zhang et al., 2020)	0.485	0.345	0.524	**0.558**	**0.533**	0.463	0.580	0.347	0.352	0.325	0.274
EED‡ (Stanchev et al., 2019)	0.431	0.315	0.508	**0.568**	0.518	0.425	0.546	0.257	0.345	0.301	0.267
ESIM‡ (Mathur et al., 2019)	–	0.329	0.511	–	0.510	0.428	0.572	0.339	0.331	0.290	0.289
YISI-1‡ (Lo, 2019)	0.475	0.351	0.537	**0.551**	**0.546**	0.470	0.585	**0.355**	0.376	0.349	0.310
YISI-1_SRL‡ (Lo, 2019)	–	0.368	–	–	–	–	–	**0.361**	–	–	0.299
Prism-ref (This Work)	**0.582**	**0.426**	**0.591**	0.313	**0.531**	**0.558**	0.584	**0.376**	**0.458**	**0.453**	**0.426**
LASER + LM (Contrastive)	0.535	0.402	0.568	0.306	0.408	0.503	**0.640**	**0.356**	0.431	0.401	**0.381**

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
BERTSCORE (Zhang et al., 2019, 2020)	0.176	**0.345**	**0.320**	**0.432**	**0.381**	**0.223**	**0.430**
EED‡ (Stanchev et al., 2019)	0.120	0.281	0.264	0.392	0.298	0.176	0.376
ESIM‡ (Chen et al., 2017; Mathur et al., 2019)	0.167	0.337	0.303	**0.435**	0.359	0.201	0.396
YISI-1‡ (Lo, 2019)	0.164	**0.347**	**0.312**	**0.440**	**0.376**	**0.217**	**0.426**
YISI-1_SRL‡ (Lo, 2019)	**0.199**	**0.346**	0.306	**0.442**	**0.380**	**0.222**	**0.431**
Prism-ref (This Work)	**0.204**	**0.357**	**0.313**	**0.434**	**0.382**	**0.225**	**0.438**
Prism-ref w/ Parabank2 (Contrastive)	0.184	**0.341**	**0.326**	**0.425**	**0.373**	0.207	**0.432**
LASER + LM (Contrastive)	0.190	0.335	**0.319**	**0.428**	**0.368**	0.207	0.416

Table 2: WMT19 segment-level human correlation (τ), to non-English (top) and to English (bottom). **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. ‡:WMT19 Metric Submission. For brevity, only competitive baselines are shown. For complete results see Appendix D. Note that our models were not trained on Gujarati (gu). "LASER + LM" denotes the optimal linear combination found on the development set.

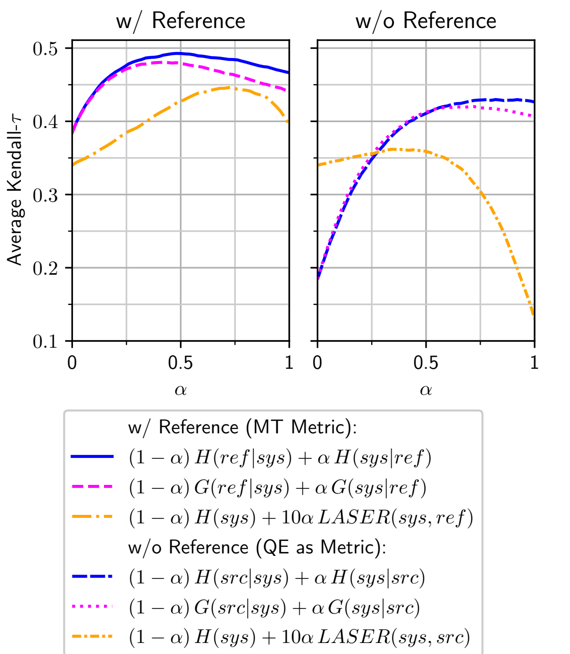

Figure 4: Linear combinations of scoring each direction using length-normalized (H) vs un-normalized (G) log probability for our method, and length-normalized language model probabilities (H) vs LASER for our contrastive method. In both cases, we explore scoring using the human reference ref vs the source src. Results are segment-level τ on our development set (WMT18), averaged across all language pairs. Full results on WMT18 are provided in Appendix C.

between LASER embeddings of sys and ref [$LASER(\text{sys}, \text{ref})$] both have decent correlation with human judgments and are complementary. However, cosine distance between LASER embeddings of sys and src [$LASER(\text{sys}, \text{src})$] has only weak correlation.[11]

5.3 Segment-Level Metric Results

Segment-level metric results are shown in Table 2. On language pairs in to non-English, we outperform prior work by a statistically significant margin in 8 of 11 language pairs and are statistically tied in the rest,[12] with the exception of Gujarati (gu) where the model had no training data. In to English, our metric is statistically tied with the best prior work in every language pair. Our metric tends to significantly outperform our contrastive LASER + LM method, although the contrastive method performs surprisingly well in en-ru.

5.4 System-Level Metric Results

Table 3 shows system-level metric performance compared to BLEU, BERTscore, and Yisi variants on the top four systems submitted to WMT19. Results for all metrics on the top four systems, with statistical significance, is provided in Appendix E.

[11]This corroborates findings of Fonseca et al. (2019).

[12]Rerunning without LASER+LM shows Prism-ref to be statistically tied with Yisi-1, ESIM, and BERTscore.

	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	de-cs	de-fr	fr-de
BERTSCORE (Zhang et al., 2020)	0.868	**−0.722**	0.859	0.922	0.288	0.955	0.953	0.982	0.976	0.707	0.973
BLEU† (Papineni et al., 2002)	0.930	**−0.370**	0.898	0.860	0.181	0.925	0.753	0.987	0.812	0.495	0.983
YISI-1‡ (Lo, 2019)	0.847	**−0.220**	0.976	0.917	0.342	0.838	0.963	0.990	0.967	0.677	0.967
YISI-1_SRL‡ (Lo, 2019)	–	**−0.378**	–	–	–	–	–	0.994	–	–	0.974
Prism-ref (This Work)	0.952	0.278	0.886	0.863	0.693	0.862	0.975	0.966	0.968	0.648	0.998
LASER + LM (Contrastive)	0.961	0.377	0.903	0.509	0.605	0.743	0.962	0.985	0.947	0.774	0.975

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
BERTSCORE (Zhang et al., 2020)	0.272	0.683	0.913	0.897	0.753	0.456	**−0.220**
BLEU† (Papineni et al., 2002)	**−0.822**	**−0.275**	0.966	0.958	0.625	**−0.356**	**−0.694**
YISI-1‡ (Lo, 2019)	0.045	0.610	0.962	0.887	0.552	0.365	**−0.067**
YISI-1_SRL‡ (Lo, 2019)	0.081	0.580	0.959	0.874	0.560	0.342	**−0.069**
Prism-ref (This Work)	0.401	0.719	0.896	0.796	0.877	0.431	0.523
LASER + LM (Contrastive)	0.957	0.768	0.867	0.870	0.615	0.596	0.733

Table 3: WMT19 system-level human correlation (Pearson), for top 4 systems only, to non-English (top) and to English (bottom). Negative correlations with human judgments shown in **red** for emphasis. †:WMT19 Baseline ‡:WMT19 Metric Submission. "LASER + LM" denotes the optimal linear combination found on the development set. Note that our models were not trained on Gujarati (gu).

While correlations with human judgments are not high in all cases for our metric, they are at least positive. In contrast, BLEU has negative correlation in 5 language pairs, and BERTscore and Yisi-1 variants are each negative in at least two.

We do not find the system-level results on all submitted MT systems (see Appendix F) to be particularly interesting; as noted by Ma et al. (2019), a single weak system can result in high overall system-level correlation, even for an otherwise poorly performing metric.

5.5 QE as a Metric Results

We find that our reference-less Prism-src outperforms all QE as a metrics systems from the WMT19 shared task, by a statistically significant margin, in every language pair at segment-level human correlation (Table 4). Prism-src also outperforms or statistically ties with every QE as a metric systems at system-level human correlation (Appendix F).

6 Analysis and Discussion

Are Human References Helpful? The fact that our model is multilingual allows us to explore the extent to which the human reference actually improves our model's ability to judge MT system output, compared to using the source instead.[13] Comparing the performance of our method with access to the human reference (Prism-ref) vs our method with access to only the source (Prism-src), we find that the reference-based method statistically outperforms the source-based method in all but one language pair. We find the case where they are not statistically different, de-cs, to be particularly interesting: de-cs was the only language pair in WMT 19 where the systems were unsupervised (i.e., did not use parallel training data). As a result, it is the only language pair where our model outperformed the best WMT system at translation. In most cases, our model is *substantially worse* at translation than the best WMT system (see Appendix G); thus the performance difference between Prism-ref and Prism-src would suggest that the model needs no help in judging MT systems which are weaker than it is, but *the human references are assisting our model in evaluating MT systems which are stronger than it is.* This means that we have not simply reduced the task of MT evaluation to that of building a SOTA MT system. We see that a reasonably good (but not SOTA) multilingual NMT system, with help from the human translator(s) that produced the references, can be a SOTA MT metric and judge SOTA MT systems.

Does our Method Exhibit Reference Bias? Human judgments of MT system output can be made using either the source (by bilingual annotators) or a human reference (by monolingual annotators). In WMT19, judgments for translations in to English were reference-based, while translations in to non-English were source-based. Fomicheva and Specia

[13]The underlying assumption in all of MT metrics is that the work done by the human translator makes it is easier to automatically judge the quality of MT output. However, if our model or the MT systems being judged were strong enough, we would expect this assumption to break down.

	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	de-cs	de-fr	fr-de
Best WMT19 QE as Metric	0.069	0.236	0.351	0.147	0.187	0.003	0.226	0.044	0.199	0.186	0.066
Prism-src (This work)	**0.470**	**0.402**	**0.555**	**0.215**	**0.507**	**0.499**	**0.486**	**0.287**	**0.444**	**0.371**	**0.316**

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
Best WMT19 QE as Metric	0.068	0.211	−0.001	0.096	0.075	0.089	0.253
Prism-src (This work)	**0.109**	**0.300**	**0.102**	**0.391**	**0.356**	**0.178**	**0.336**

Table 4: WMT19 segment-level human correlation (τ) for QE as Metric systems (which have access to the source only, not the reference). **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. Note that our models were not trained on Gujarati (gu). For brevity, only the best system from WMT19 for each language pair is shown - for complete results see Appendix F.

(2016) noted that reference-based annotations unfairly reward MT system output which is similar to the reference over equally valid output which is less similar to the reference, denoted "reference-bias."[14] It is likely that MT metrics exhibit a similar bias, so we were curious if we could observe any differences in trends between language pairs with reference-based vs source-based annotations. With the exception of de-cs (discussed above), we see statistically significant improvements for Prism-ref over Prism-src (which does not use the reference, so cannot be biased by it), both in to English and in to non-English. Thus we conclude that if our method suffers from reference bias, its effects are small compared to the benefit of using the human translation.

Unbiased vs. Generative Paraphraser Our unbiased paraphraser statistically outperforms the generative English-only Parabank2 paraphraser in 6 of 7 language pairs, however wins are statistically significant in only 2 languages pairs with statistical ties in the rest. We believe this is be due to the Parabank2 model having a lexical/syntactic bias away from its input – see § 5.1 Additionally, creating synthetic paraphrases and training individual models in many languages would be a substantial undertaking.

Paraphrasing vs LASER + LM The proposed method significantly outperforms the contrastive LASER-based method in most language pairs, even when LASER is augmented with a language model. This suggests that training a multilingual paraphraser is a better use of multilingual bitext than training a sentence embedder, although the comparison is complicated by the fact that LASER is trained on different data from our model. This is consistent with neural MT showing significant improvements over statistical MT, where a phrase table and language model were trained separately and combined at decode time.

[14]This claim is disputed (Ma et al., 2017a).

7 Conclusion and Future Work

In this work, we show that a multilingual NMT system can be used as an unbiased, multilingual paraphraser, and we show that the resulting paraphraser can be used as an MT metric. Our single model supports 39 languages and outperforms prior metrics on the most recent WMT metrics shared task evaluation. We present analysis showing our method's high human judgment correlation is not simply the result of reference bias. We also present analysis showing that we have not simply reduced the task of evalution to that of building a SOTA MT system; the work done by the human translator helps the evaluation model judge systems that are stronger (at translation) than it is, and we do not need a SOTA multilingual NMT model to score SOTA MT systems or be a SOTA MT metric.

Our method outperforms metrics using highly optimized BERT variants, and we are optimistic our method will improve further as stronger multilingual NMT models become publicly available.

In future work, we would like to explore whether the unbiased paraphraser presented in this work is well suited to other other tasks, such as data augmentation. We would also like to extend this work to paragraph- or document-level evaluation by training a paragraph- or document-level multilingual NMT system, as there is growing evidence that MT evaluation would be better conducted at the document level, rather than the sentence level (Läubli et al., 2018).

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

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.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. WMDO: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 494–500, Florence, Italy. Association for Computational Linguistics.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2010. Extending the METEOR machine translation evaluation metric to the phrase level. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 250–253, Los Angeles, California. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington. 2002a. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT 02, page 138145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada. Association for Computational Linguistics.

Melania Duma and Wolfgang Menzel. 2017. UHH submission to the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 582–588, Copenhagen, Denmark. Association for Computational Linguistics.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *LREC*.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria. Association for Computational Linguistics.

Christian Federmann, Oussama Elachqar, and Chris Quirk. 2019. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 17–26, Hong Kong, China. Association for Computational Linguistics.

Marina Fomicheva and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 77–82, Berlin, Germany. Association for Computational Linguistics.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 4276–4283, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Yinuo Guo and Junfeng Hu. 2019. Meteor++ 2.0: Adopt syntactic level paraphrase knowledge into machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 501–506, Florence, Italy. Association for Computational Linguistics.

Rohit Gupta, Constantin Orăsan, and Josef van Genabith. 2015. ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal. Association for Computational Linguistics.

Aaron L.-F Han, Derek Wong, Lidia Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013. Mt summit13.language-independent model for machine translation evaluation with reinforced factors.

Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. 2012. LEPOR: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING 2012: Posters*, pages 441–450, Mumbai, India. The COLING 2012 Organizing Committee.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming

Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019a. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019b. ParaBank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of AAAI*.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019c. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, and zhifeng Chen. 2019. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 103–112. Curran Associates, Inc.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. deepQuest: A framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Rakesh Khobragade, Heaven Patel, Anand Namdev, Anish Mishra, and Pushpak Bhattacharyya. 2019. Machine translation evaluation using bi-directional entailment.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597, Copenhagen, Denmark. Association for Computational Linguistics.

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

Chi-kiu Lo and Dekai Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 220–229, Portland, Oregon, USA. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Yvette Graham, Timothy Baldwin, and Qun Liu. 2017a. Further investigation into reference bias in monolingual evaluation of machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2476–2485, Copenhagen, Denmark. Association for Computational Linguistics.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017b. Blend: a novel combined MT metric based on direct assessment — CASICT-DCU submission to WMT17 metrics task. In *Proceedings of the Second Conference on Machine Translation*, pages 598–603, Copenhagen, Denmark. 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.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. *arXiv preprint arXiv:1803.00047*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 297–305, Suntec, Singapore. Association for Computational Linguistics.

Joybrata Panja and Sudip Kumar Naskar. 2018. ITER: Improving translation edit rate through optimizable edit costs. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 746–750, Belgium, Brussels. 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.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. 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.

Maja Popović, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: IBM1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 99–103, Edinburgh, Scotland. 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.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2923–2934, Osaka, Japan. The COLING 2016 Organizing Committee.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 142–149, Barcelona, Spain. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Alessandro Raganato, Raúl Vázquez, Mathias Creutz, and Jörg Tiedemann. 2019. An evaluation of language-agnostic inner-attention-based representations in machine translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 27–32, Florence, Italy. Association for Computational Linguistics.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

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 and Khalil Sima'an. 2015. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 396–401, Lisbon, Portugal. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Jörg Tiedemann and Yves Scherrer. 2019. Measuring semantic abstraction of multilingual NMT with paraphrase recognition and generation tasks. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 35–42, Minneapolis, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4602–4608, Florence, Italy. Association for Computational Linguistics.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285, Copenhagen, Denmark. Association for Computational Linguistics.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 101–105, Florence, Italy. Association for Computational Linguistics.

Ryoma Yoshimura, Hiroki Shimanaka, Yukio Matsumura, Hayahide Yamagishi, and Mamoru Komachi. 2019. Filtering pseudo-references by paraphrasing for automatic evaluation of machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 521–525, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. ParaEval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 447–454, New York City, USA. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Generation Examples

Figure 5 shows sentences generated from both our model and parabank2.

REFERENCE	28-Year-Old Chef Found Dead at San Francisco Mall
THIS WORK	28-Year-Old Chef Found Dead at San Francisco Mall
PARABANK2	28-year-**o**ld **c**hef **f**ound **d**ead **in a mall in** San Francisco
REFERENCE	A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
THIS WORK	A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
PARABANK2	**Earlier this week,** a 28-year-old chef who ~~had~~ recently moved to San Francisco was found dead **on** the **steps** of a local **department store**.
REFERENCE	But the victim's brother says he can't think of anyone who would want to hurt him, saying, "Things were finally going well for him."
THIS WORK	But the victim's brother says he can't think of anyone who would want to hurt him, saying, "Things were finally going well for him."
PARABANK2	But the victim's brother said he **couldn't** think of anyone who**'d** want to hurt him, **and he said he was finally okay.**
REFERENCE	The body found at the Westfield Mall Wednesday morning was identified as 28-year-old San Francisco resident Frank Galicia, the San Francisco Medical Examiner's Office said.
THIS WORK	The body found at ~~the~~ Westfield Mall Wednesday morning was identified as 28-year-old San Francisco resident Frank Galicia, the San Francisco Medical Examiner's Office said.
PARABANK2	The body found **Wednesday morning** at the Westfield Mall **has been** identified **by the San Francisco Medical Examiner's Office** as 28-year-old ~~**San Francisco resident**~~ Frank Galicia.
REFERENCE	The San Francisco Police Department said the death was ruled a homicide and an investigation is ongoing.
THIS WORK	The San Francisco Police Department said the death was **deemed** a homicide and an investigation is ongoing.
PARABANK2	~~The~~ San Francisco **P.D. says** the death **has been ruled** a **murder** and **is under investigation**.
REFERENCE	The victim's brother, Louis Galicia, told ABC station KGO in San Francisco that Frank, previously a line cook in Boston, had landed his dream job as line chef at San Francisco's Sons & Daughters restaurant six months ago.
THIS WORK	The victim's brother, Louis Galicia, told ABC station KGO in San Francisco that Frank, formerly a line cook in Boston, had landed his dream job as line chef at San Francisco's Sons & Daughters restaurant six months ago.
PARABANK2	~~The~~ Victim's brother, Louis Galicia, told ABC station KGO in San Francisco that Frank, **who used to be** a line chef in Boston, **quit** his dream job **six months ago** as a line chef at the Sons & Daughters Restaurant in San Francisco.
REFERENCE	A spokesperson for Sons & Daughters said they were "shocked and devastated" by his death.
THIS WORK	A spokesperson for Sons & Daughters said they were "shocked and devastated" by his death
PARABANK2	A **spokesman** for Sons & Daughters said that **his death** "shocked and devastated **them.**"
REFERENCE	"We are a small team that operates like a close knit family and he will be dearly missed," the spokesperson said.
THIS WORK	"We are a small team that operates like a close-knit family and he will be dearly missed," the **spokesman** said.
PARABANK2	"We are a small team, **operating** as a close-knit family**,** and **we will miss him** dearly," said the **spokesman**.
REFERENCE	Our thoughts and condolences are with Frank's family and friends at this difficult time.
THIS WORK	Our thoughts and condolences are with Frank's family and friends at this difficult time.
PARABANK2	Our thoughts and condolences go out to Frank's family and friends in these difficult times.
REFERENCE	Louis Galicia said Frank initially stayed in hostels, but recently, "Things were finally going well for him."
THIS WORK	Louis Galicia said Frank initially stayed in hostels, but recently, "Things were finally going well for him."
PARABANK2	Louis Galicia said **that** Frank initially stayed in **the dormitory**, but **lately**, "**He's finally doing okay**."

Figure 5: Sentences generated via beam search (beamwidth 5) for the multilingual model presented in this work vs parabank2. We note that our model tends to produces copies or near copies of the input, which is the desired behavior for our application. Changes are emphasized with **bold** or ~~**strikethrough**~~. The parabank2 model tends to produce output with lexical/syntactic changes, which occasionally also significantly change the meaning of the sentence (denoted in red). References (paraphraser inputs) are the first ten sentences of wmt17 zh-en.

B Data Details for Replication

The bulk of our data comes from Wikimatrix (Schwenk et al., 2019), a large collection of parallel data extracted from Wikipedia, and for more domain variety, we added Global Voices,[15] EuroParl (Koehn, 2005) (random subset of to 100k sentence pairs per language pair), SETimes,[16] United Nations (Eisele and Chen, 2010) (random sample of 1M sentence pairs per language pair). We also included WMT Kazakh-English and Kazakh-Russian data from WMT, to be able to evaluate on Kazakh.

WMT Kazakh-English and Kazakh-Russian were limited to the best 1M and 200k sentence pairs, respectively, as judged by LASER. We used a margin threshold of 1.05 for Wikimatrix and a threshold of 1.04 for the remaining datasets, as we expect them to be cleaner. We find that FastText classifies many sentences as non-English when they contain mostly English but also contain a few non-English words, especially from lower resource languages. To remedy this, we performed LID on 5-grams and filtered out sentences for which LID did not classify at least half of the 5-grams as the expected language.

We filtered out sentences where there was more than 60% overlap in 3-grams or 40% overlap in 4-grams.Via manual inspection, this seemed to provide a good trade-off between allowing numbers and named entities to be copied, but filtering out sentences that were clearly not translated. We perform tokenization with sentencepiece prior to filtering, using a 200k vocabulary for all language pairs, to account for languages like Chinese which do not denote word boundaries. Note that this vocabulary was used only for filtering, not for training the final model.

We limited training to languages with at least 1M examples, which resulted in 39 languages.

Figure 6 shows amount of data in each language.

[15]http://casmacat.eu/corpus/global-voices.html

[16]http://nlp.ffzg.hr/resources/corpora/setimes/

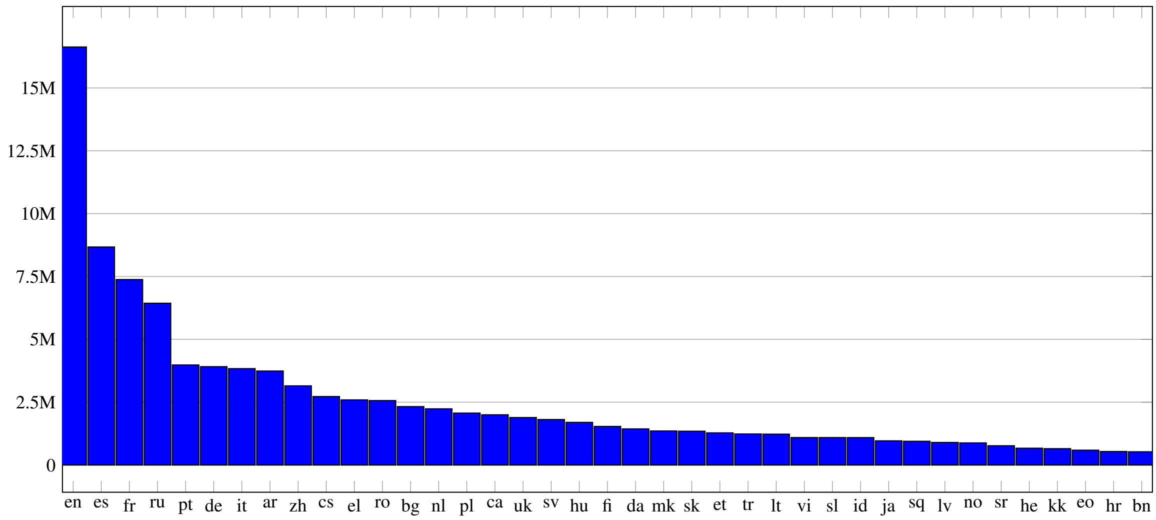

Figure 6: Distribution of the 39 languages (ISO 639-1 language code) of the 99.8M training sentences. English accounts for 16.7%. Spanish, French, Russian, Portuguese, German, and Italian account for a combined 34.3%. The bottom 20 languages account for only 21.9% combined.

C WMT 2018 System- and Segment-Level Results

Table 5, Table 6, Table 7, and Table 8 show system- and segment- level results, in to and out of English, for the WMT 2018 MT metrics shared task, along with all baselines and submitted systems.

n	**en-cs** 5413	**en-de** 19711	**en-et** 32202	**en-fi** 9809	**en-ru** 22181	**en-tr** 1358	**en-zh** 28602
BEER‡ (Stanojević and Sima'an, 2015)	0.518	0.686	0.558	0.511	0.403	0.374	0.302
BERTSCORE (Zhang et al., 2019, 2020)	0.559	0.727	0.584	0.538	0.424	0.389	**0.364**
BLEND‡ (Ma et al., 2017b)	–	–	–	–	0.394	–	–
CHARACTER‡ (Wang et al., 2016)	0.414	0.604	0.464	0.403	0.352	0.404	0.313
CHRF† (Popović, 2015)	0.516	0.677	0.572	0.520	0.383	0.409	0.328
CHRF+† (Popović, 2017)	0.513	0.680	0.573	0.525	0.392	0.405	0.328
ITER‡ (Panja and Naskar, 2018)	0.333	0.610	0.392	0.311	0.291	0.236	–
SENTBLEU† (Papineni et al., 2002)	0.389	0.620	0.414	0.355	0.330	0.261	0.311
YISI-0‡ (Lo, 2019)	0.471	0.661	0.531	0.464	0.394	0.376	0.318
YISI-1‡ (Lo, 2019)	0.496	0.691	0.546	0.504	0.407	0.418	0.323
YISI-1_SRL‡ (Lo, 2019)	–	0.696	–	–	–	–	0.310
Prism-ref (This Work)	**0.667**	**0.799**	**0.705**	**0.667**	**0.469**	**0.574**	**0.371**
LASER + LM (Contrastive)	0.587	0.746	0.628	0.629	**0.450**	**0.501**	**0.367**
Prism-src (This work)	0.552	0.732	0.636	0.626	0.409	**0.505**	0.298
LM	0.459	0.655	0.408	0.511	0.375	0.331	0.221
LASER	0.480	0.677	0.585	0.511	0.402	0.432	0.338

Table 5: WMT18 Segment-level results, from English. n denotes number of pairwise judgments. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT18 Baseline (Ma et al., 2018) ‡:WMT18 Metric Submission (Ma et al., 2018)

n	**cs-en** 5110	**de-en** 77811	**et-en** 56721	**fi-en** 15648	**ru-en** 10404	**tr-en** 8525	**zh-en** 33357
BEER‡ (Stanojević and Sima'an, 2015)	0.295	0.481	0.341	0.232	0.288	0.229	0.214
BERTSCORE (Zhang et al., 2019, 2020)	**0.404**	0.550	**0.397**	**0.296**	**0.340**	**0.292**	**0.253**
BLEND‡ (Ma et al., 2017b)	0.322	0.492	0.354	0.226	0.290	0.232	0.217
CHARACTER‡ (Wang et al., 2016)	0.256	0.450	0.286	0.185	0.244	0.172	0.202
CHRF† (Popović, 2015)	0.288	0.479	0.328	0.229	0.269	0.210	0.208
CHRF+† (Popović, 2017)	0.288	0.479	0.332	0.234	0.279	0.218	0.207
ITER‡ (Panja and Naskar, 2018)	0.198	0.396	0.235	0.128	0.139	-0.029	0.144
METEOR++‡ (Shimanaka et al., 2018)	0.270	0.457	0.329	0.207	0.253	0.204	0.179
RUSE‡ (Shimanaka et al., 2018)	0.347	0.498	0.368	0.273	0.311	0.259	0.218
SENTBLEU† (Papineni et al., 2002)	0.233	0.415	0.285	0.154	0.228	0.145	0.178
UHH_TSKM‡ (Duma and Menzel, 2017)	0.274	0.436	0.300	0.168	0.235	0.154	0.151
YISI-0‡ (Lo, 2019)	0.301	0.474	0.330	0.225	0.294	0.215	0.205
YISI-1‡ (Lo, 2019)	0.319	0.488	0.351	0.231	0.300	0.234	0.211
YISI-1_SRL‡ (Lo, 2019)	0.317	0.483	0.345	0.237	0.306	0.233	0.209
Prism-ref (This Work)	**0.423**	**0.560**	**0.409**	**0.317**	**0.366**	**0.309**	**0.263**
Prism-ref w/ Parabank2 (Contrastive)	**0.386**	0.538	**0.399**	**0.309**	**0.340**	**0.275**	0.244
LASER + LM (Contrastive)	0.364	0.526	0.378	0.265	0.305	0.257	0.243
Prism-src (This work)	0.355	0.515	0.370	0.257	0.308	0.213	0.194
LM	0.285	0.438	0.285	0.198	0.280	0.123	0.192
LASER	0.310	0.494	0.364	0.232	0.257	0.248	0.207

Table 6: WMT18 Segment-level results, to English. n denotes number of pairwise judgments. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT18 Baseline (Ma et al., 2018) ‡:WMT18 Metric Submission (Ma et al., 2018)

n	en-cs 5	en-de 16	en-et 14	en-fi 12	en-ru 9	en-tr 8	en-zh 14
BEER[‡] (Stanojević and Sima'an, 2015)	**0.992**	**0.991**	**0.980**	**0.961**	**0.988**	**0.965**	0.928
BERTSCORE (Zhang et al., 2019, 2020)	**0.997**	**0.989**	**0.982**	**0.972**	**0.990**	0.908	0.967
BLEND[‡] (Ma et al., 2017b)	–	–	–	–	**0.988**	–	–
BLEU[†] (Papineni et al., 2002)	0.995	0.981	0.975	**0.962**	0.983	0.826	0.947
CDER[†] (Leusch et al., 2006)	0.997	0.986	**0.984**	**0.964**	**0.984**	0.861	0.961
CHARACTER[‡] (Wang et al., 2016)	**0.993**	**0.989**	0.956	**0.974**	0.983	0.833	**0.983**
CHRF[†] (Popović, 2015)	**0.990**	**0.990**	**0.981**	**0.969**	**0.989**	0.948	0.944
CHRF+[†] (Popović, 2017)	**0.990**	0.989	**0.982**	**0.970**	0.989	0.943	0.943
ITER[‡] (Panja and Naskar, 2018)	0.915	**0.984**	**0.981**	**0.973**	0.975	0.865	–
NIST[†] (Doddington, 2002b)	**0.999**	0.986	**0.983**	0.949	**0.990**	0.902	0.950
PER[†]	0.991	0.981	0.958	0.906	**0.988**	0.859	0.964
TER[†] (Snover et al., 2006)	**0.997**	**0.988**	**0.981**	0.942	**0.987**	0.867	**0.963**
WER[†]	**0.997**	**0.986**	**0.981**	0.945	0.985	0.853	0.957
YISI-0[‡] (Lo, 2019)	0.973	0.985	0.968	0.944	**0.990**	**0.990**	0.957
YISI-1[‡] (Lo, 2019)	**0.987**	0.985	**0.979**	0.940	**0.992**	**0.976**	0.963
YISI-1_SRL[‡] (Lo, 2019)	–	**0.990**	–	–	–	–	0.952
Prism-ref (This Work)	0.962	**0.987**	**0.973**	**0.976**	**0.989**	0.894	**0.977**
LASER + LM (Contrastive)	0.953	**0.984**	**0.980**	**0.976**	0.984	0.927	**0.982**
Prism-src (This work)	0.850	**0.984**	0.949	0.964	0.960	0.864	0.940
LM	0.854	**0.985**	0.837	0.938	0.959	0.830	0.859
LASER	**0.995**	0.965	0.937	**0.978**	**0.993**	0.895	**0.978**

Table 7: WMT18 System-level results, from English. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT18 Baseline (Ma et al., 2018) ‡:WMT18 Metric Submission (Ma et al., 2018)

n	cs-en 5	de-en 16	et-en 14	fi-en 9	ru-en 8	tr-en 5	zh-en 14
BEER[‡] (Stanojević and Sima'an, 2015)	0.958	0.994	**0.985**	0.991	0.982	0.870	**0.976**
BERTSCORE (Zhang et al., 2019, 2020)	**0.990**	**0.999**	**0.990**	**0.998**	0.935	0.499	0.956
BLEND[‡] (Ma et al., 2017b)	0.973	0.991	0.985	**0.994**	**0.993**	**0.801**	**0.976**
BLEU[†] (Papineni et al., 2002)	0.970	0.971	**0.986**	0.973	0.979	**0.657**	**0.978**
CDER[†] (Leusch et al., 2006)	0.972	0.980	**0.990**	0.984	0.980	**0.664**	**0.982**
CHARACTER[‡] (Wang et al., 2016)	**0.970**	0.993	0.979	0.989	**0.991**	**0.782**	0.950
CHRF[†] (Popović, 2015)	0.966	0.994	0.981	0.987	**0.990**	0.452	0.960
CHRF+[†] (Popović, 2017)	0.966	0.993	0.981	0.989	0.990	0.174	0.964
ITER[‡] (Panja and Naskar, 2018)	0.975	0.990	0.975	**0.996**	0.937	**0.861**	**0.980**
METEOR++[‡] (Shimanaka et al., 2018)	0.945	0.991	0.978	0.971	**0.995**	0.864	0.962
NIST[†] (Doddington, 2002b)	0.954	0.984	0.983	0.975	0.973	**0.970**	0.968
PER[†]	**0.970**	0.985	**0.983**	0.993	0.967	0.159	0.931
RUSE[‡] (Shimanaka et al., 2018)	0.981	0.997	**0.990**	0.991	**0.988**	**0.853**	**0.981**
TER[†] (Snover et al., 2006)	**0.950**	0.970	**0.990**	0.968	0.970	0.533	0.975
UHH_TSKM[‡] (Duma and Menzel, 2017)	0.952	0.980	**0.989**	0.982	0.980	0.547	**0.981**
WER[†]	**0.951**	0.961	**0.991**	0.961	0.968	0.041	0.975
YISI-0[‡] (Lo, 2019)	0.956	0.994	0.975	0.978	0.988	**0.954**	0.957
YISI-1[‡] (Lo, 2019)	0.950	0.992	0.979	0.973	**0.991**	**0.958**	0.951
YISI-1_SRL[‡] (Lo, 2019)	0.965	0.995	0.981	0.977	**0.992**	**0.869**	0.962
Prism-ref (This Work)	0.988	0.995	0.971	**0.998**	**0.995**	0.730	**0.989**
Prism-ref w/ Parabank2 (Contrastive)	**0.992**	0.989	0.964	**0.998**	**0.996**	**0.896**	0.986
LASER + LM (Contrastive)	**0.988**	0.991	0.965	0.994	0.745	0.297	0.890
Prism-src (This work)	**0.984**	0.991	0.964	0.987	0.970	**0.896**	0.958
LM	**0.986**	0.970	0.954	0.898	0.951	**0.891**	0.972
LASER	**0.978**	0.986	0.953	0.984	0.489	**0.968**	0.591

Table 8: WMT18 System-level results, to English. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT18 Baseline (Ma et al., 2018) ‡:WMT18 Metric Submission (Ma et al., 2018)

D WMT 2019 Metric and QE as Metric Segment-Level Results

Table 9, Table 10, and Table 11 show segment-level metrics (excluding QE as a metric) results, for language pairs in to, out of, and not including English, for the WMT 2019 MT metrics shared task, along with all baselines and submitted systems.

Table 12, Table 13, and Table 14 show segment-level QE as a metric results, for language pairs in to, out of, and not including English, for the WMT 2019 MT metrics shared task, along with all baselines and submitted systems.

n	**en-cs** 27178	**en-de** 99840	**en-fi** 31820	**en-gu** 11355	**en-kk** 18172	**en-lt** 17401	**en-ru** 24334	**en-zh** 18658
BEER‡ (Stanojević and Sima'an, 2015)	0.443	0.316	0.514	0.537	0.516	0.441	0.542	0.232
BERTScore (Zhang et al., 2019, 2020)	0.485	0.345	0.524	**0.558**	**0.533**	0.463	0.580	0.347
CharacTER‡ (Wang et al., 2016)	0.349	0.264	0.404	0.500	0.351	0.311	0.432	0.094
chrF† (Popović, 2015)	0.455	0.326	0.514	0.534	0.479	0.446	0.539	0.301
chrF+† (Popović, 2017)	0.458	0.327	0.514	0.538	0.491	0.448	0.543	0.296
EED‡ (Stanchev et al., 2019)	0.431	0.315	0.508	**0.568**	0.518	0.425	0.546	0.257
ESIM‡ (Chen et al., 2017; Mathur et al., 2019)	–	0.329	0.511	–	0.510	0.428	0.572	0.339
hLEPORA_baseline‡ (Han et al., 2012, 2013)	–	–	–	0.463	0.390	–	–	–
sentBLEU† (Papineni et al., 2002)	0.367	0.248	0.396	0.465	0.392	0.334	0.469	0.270
YiSi-0‡ (Lo, 2019)	0.406	0.304	0.483	0.539	0.494	0.402	0.535	0.266
YiSi-1‡ (Lo, 2019)	0.475	0.351	0.537	**0.551**	**0.546**	0.470	0.585	**0.355**
YiSi-1_srl‡ (Lo, 2019)	–	0.368	–	–	–	–	–	**0.361**
Prism-ref (This Work)	**0.582**	**0.426**	**0.591**	0.313	**0.531**	**0.558**	0.584	**0.376**
LASER + LM (Contrastive)	0.535	0.402	0.568	0.306	0.408	0.503	**0.640**	**0.356**

Table 9: WMT19 Segment-level results, metrics (excludes QE as metric results), from English. n denotes number of pairwise judgments. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT19 Baseline (Ma et al., 2019) ‡:WMT19 Metric Submission (Ma et al., 2019)

n	**de-en** 85365	**fi-en** 38307	**gu-en** 31139	**kk-en** 27094	**lt-en** 21862	**ru-en** 46172	**zh-en** 31070
BEER‡ (Stanojević and Sima'an, 2015)	0.128	0.283	0.260	0.421	0.315	0.189	0.371
BERTr‡ (Mathur et al., 2019)	0.142	0.331	0.291	0.421	0.353	0.195	0.399
BERTScore (Zhang et al., 2019, 2020)	0.176	**0.345**	**0.320**	**0.432**	**0.381**	**0.223**	**0.430**
CharacTER‡ (Wang et al., 2016)	0.101	0.253	0.190	0.340	0.254	0.155	0.337
chrF† (Popović, 2015)	0.122	0.286	0.256	0.389	0.301	0.180	0.371
chrF+† (Popović, 2017)	0.125	0.289	0.257	0.394	0.303	0.182	0.374
EED‡ (Stanchev et al., 2019)	0.120	0.281	0.264	0.392	0.298	0.176	0.376
ESIM‡ (Chen et al., 2017; Mathur et al., 2019)	0.167	0.337	0.303	**0.435**	0.359	0.201	0.396
hLEPORA_baseline‡ (Han et al., 2012, 2013)	–	–	–	0.372	–	–	0.339
Meteor++_2.0(syntax)‡ (Guo and Hu, 2019)	0.084	0.274	0.237	0.395	0.291	0.156	0.370
Meteor++_2.0(syntax+copy)‡ (Guo and Hu, 2019)	0.094	0.273	0.244	0.402	0.287	0.163	0.367
PReP‡ (Yoshimura et al., 2019)	0.030	0.197	0.192	0.386	0.193	0.124	0.267
sentBLEU† (Papineni et al., 2002)	0.056	0.233	0.188	0.377	0.262	0.125	0.323
WMDO‡ (Chow et al., 2019)	0.096	0.281	0.260	0.420	0.300	0.162	0.362
YiSi-0‡ (Lo, 2019)	0.117	0.271	0.263	0.402	0.289	0.178	0.355
YiSi-1‡ (Lo, 2019)	0.164	**0.347**	**0.312**	**0.440**	**0.376**	**0.217**	**0.426**
YiSi-1_srl‡ (Lo, 2019)	**0.199**	**0.346**	0.306	**0.442**	**0.380**	**0.222**	**0.431**
Prism-ref (This Work)	**0.204**	**0.357**	**0.313**	**0.434**	**0.382**	**0.225**	**0.438**
Prism-ref w/ Parabank2 (Contrastive)	0.184	**0.341**	**0.326**	**0.425**	**0.373**	0.207	**0.432**

Table 10: WMT19 Segment-level results, metrics (excludes QE as metric), to English. n denotes number of pairwise judgments. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT19 Baseline (Ma et al., 2019) ‡:WMT19 Metric Submission (Ma et al., 2019)

n	de-cs 35793	de-fr 4862	fr-de 1369
BEER‡ (Stanojević and Sima'an, 2015)	0.337	0.293	0.265
BERTSCORE (Zhang et al., 2019, 2020)	0.352	0.325	0.274
CHARACTER‡ (Wang et al., 2016)	0.232	0.251	0.224
CHRF† (Popović, 2015)	0.326	0.284	0.275
CHRF+† (Popović, 2017)	0.326	0.284	0.278
EED‡ (Stanchev et al., 2019)	0.345	0.301	0.267
ESIM‡ (Chen et al., 2017; Mathur et al., 2019)	0.331	0.290	0.289
HLEPORA_BASELINE‡ (Han et al., 2012, 2013)	0.207	0.239	–
SENTBLEU† (Papineni et al., 2002)	0.203	0.235	0.179
YISI-0‡ (Lo, 2019)	0.331	0.296	0.277
YISI-1‡ (Lo, 2019)	0.376	0.349	0.310
YISI-1_SRL‡ (Lo, 2019)	–	–	0.299
Prism-ref (This Work)	**0.458**	**0.453**	**0.426**
LASER + LM (Contrastive)	0.431	0.401	**0.381**

Table 11: WMT19 Segment-level results, metrics (excludes QE as metric), non-English. n denotes number of pairwise judgments. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT19 Baseline (Ma et al., 2019) ‡:WMT19 Metric Submission (Ma et al., 2019)

n	en-cs 27178	en-de 99840	en-fi 31820	en-gu 11355	en-kk 18172	en-lt 17401	en-ru 24334	en-zh 18658
IBM1-MORPHEME* (Popović et al., 2011)	−0.135	−0.003	−0.005	–	–	−0.165	–	–
IBM1-POS4GRAM* (Popović et al., 2011)	–	−0.123	–	–	–	–	–	–
LASIM*	–	0.147	–	–	–	–	−0.24	–
LP*	–	−0.119	–	–	–	–	−0.158	–
UNI* (Yankovskaya et al., 2019)	0.060	0.129	0.351	–	–	–	0.226	–
UNI+* (Yankovskaya et al., 2019)	–	–	–	–	–	–	0.222	–
USFD* (Ive et al., 2018)	–	−0.029	–	–	–	–	0.136	–
USFD-TL* (Ive et al., 2018)	–	−0.037	–	–	–	–	0.191	–
YISI-2* (Lo, 2019)	0.069	0.212	0.239	0.147	0.187	0.003	−0.155	0.044
YISI-2_SRL* (Lo, 2019)	–	0.236	–	–	–	–	–	0.034
Prism-src (This work)	**0.470**	**0.402**	**0.555**	**0.215**	**0.507**	**0.499**	**0.486**	**0.287**

Table 12: WMT19 Segment-level results, QE as a metric, from English. n denotes number of pairwise judgments. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. ∗:WMT19 QE-as-Metric Submission (Fonseca et al., 2019)

n	de-en 85365	fi-en 38307	gu-en 31139	kk-en 27094	lt-en 21862	ru-en 46172	zh-en 31070
IBM1-MORPHEME* (Popović et al., 2011)	−0.074	0.009	–	–	0.069	–	–
IBM1-POS4GRAM* (Popović et al., 2011)	−0.153	–	–	–	–	–	–
LASIM*	−0.024	–	–	–	–	0.022	–
LP*	−0.096	–	–	–	–	−0.035	–
UNI* (Yankovskaya et al., 2019)	0.022	0.202	–	–	–	0.084	–
UNI+* (Yankovskaya et al., 2019)	0.015	0.211	–	–	–	0.089	–
YISI-2* (Lo, 2019)	0.068	0.126	−0.001	0.096	0.075	0.053	0.253
YISI-2_SRL* (Lo, 2019)	0.068	–	–	–	–	–	0.246
Prism-src (This work)	**0.109**	**0.300**	**0.102**	**0.391**	**0.356**	**0.178**	**0.336**

Table 13: WMT19 Segment-level results, QE as a metric, to English. n denotes number of pairwise judgments. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. ∗:WMT19 QE-as-Metric Submission (Fonseca et al., 2019)

	de-cs	de-fr	fr-de
n	35793	4862	1369
IBM1-MORPHEME* (Popović et al., 2011)	0.048	−0.013	−0.053
IBM1-POS4GRAM* (Popović et al., 2011)	–	−0.074	−0.097
YISI-2* (Lo, 2019)	0.199	0.186	0.066
Prism-src (This work)	**0.444**	**0.371**	**0.316**

Table 14: WMT19 Segment-level results, QE as a metric, non-English. n denotes number of pairwise judgments. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. ∗:WMT19 QE-as-Metric Submission (Fonseca et al., 2019)

E WMT 2019 System-Level results for Top 4 Systems

Table 15, Table 16, and Table 17 show system-level results for just the top 4 systems, for WMT 2019. We show statistical significance following the shared task but note it appears extremely noisy.

n	en-cs 4	en-de 4	en-fi 4	en-gu 4	en-kk 4	en-lt 4	en-ru 4	en-zh 4
BEER‡ (Stanojević and Sima'an, 2015)	0.872	**−0.801**	**0.960**	0.899	0.226	**0.888**	**0.961**	**0.992**
BERTSCORE (Zhang et al., 2019, 2020)	**0.868**	−0.722	**0.859**	**0.922**	0.288	**0.955**	**0.953**	0.982
BLEU† (Papineni et al., 2002)	0.930	**−0.37**	0.898	**0.860**	0.181	**0.925**	0.753	0.987
CDER† (Leusch et al., 2006)	**0.946**	**−0.975**	0.837	**0.900**	**−0.011**	**0.880**	**0.917**	**0.986**
CHARACTER‡ (Wang et al., 2016)	0.828	**−0.777**	**0.887**	**0.902**	0.295	**0.675**	**0.974**	**0.997**
CHRF† (Popović, 2015)	0.799	−0.59	**0.936**	**0.926**	0.277	**0.901**	**0.954**	0.987
CHRF+† (Popović, 2017)	0.816	−0.605	**0.921**	**0.923**	0.283	**0.858**	0.940	**0.996**
EED‡ (Stanchev et al., 2019)	0.825	−0.552	**0.939**	**0.913**	0.267	**0.921**	0.961	**0.997**
ESIM‡ (Chen et al., 2017; Mathur et al., 2019)	–	**−0.796**	**0.957**	–	0.418	**0.997**	**0.986**	0.987
HLEPORA_BASELINE‡ (Han et al., 2012, 2013)	–	–	–	0.915	**0.062**	–	–	–
HLEPORB_BASELINE‡ (Han et al., 2012, 2013)	–	–	–	**0.915**	**0.062**	0.821	–	–
NIST† (Doddington, 2002b)	**0.946**	**−0.233**	**0.971**	**0.893**	0.082	**0.988**	0.724	0.979
PER†	0.916	**−0.995**	**0.850**	**0.887**	**−0.26**	0.390	**0.911**	**0.980**
SACREBLEU.BLEU† (Post, 2018)	**0.970**	**−0.976**	0.845	0.859	0.181	0.638	**0.878**	0.962
SACREBLEU.CHRF† (Post, 2018)	0.907	**−0.816**	**0.921**	**0.902**	0.239	**0.980**	**0.970**	**0.963**
TER† (Snover et al., 2006)	**0.969**	**−0.989**	**0.889**	0.874	**−0.06**	**0.988**	**0.895**	0.984
WER†	**0.973**	**−0.993**	0.876	0.868	**−0.058**	**0.973**	**0.894**	0.987
YISI-0‡ (Lo, 2019)	0.879	−0.796	**0.975**	**0.920**	0.196	0.787	**0.940**	0.982
YISI-1‡ (Lo, 2019)	0.847	−0.220	**0.976**	**0.917**	0.342	**0.838**	**0.963**	**0.990**
YISI-1_SRL‡ (Lo, 2019)	–	−0.378	–	–	–	–	–	**0.994**
IBM1-MORPHEME* (Popović et al., 2011)	**−0.771**	**−0.425**	**0.430**	–	–	**0.969**	–	–
IBM1-POS4GRAM* (Popović et al., 2011)	–	−0.502	–	–	–	–	–	–
LASIM*	–	−0.914	–	–	–	–	0.223	–
LP.1*	–	0.949	–	–	–	–	−0.407	–
UNI* (Yankovskaya et al., 2019)	0.587	**−0.96**	0.637	–	–	–	**0.655**	–
UNI+* (Yankovskaya et al., 2019)	–	–	–	–	–	–	0.644	–
USFD* (Ive et al., 2018)	–	**−0.729**	–	–	–	–	**0.985**	–
USFD-TL* (Ive et al., 2018)	–	−0.39	–	–	–	–	0.698	–
YISI-2* (Lo, 2019)	**0.793**	−0.933	**−0.991**	−0.389	**0.851**	−0.504	0.075	0.983
YISI-2_SRL* (Lo, 2019)	–	−0.915	–	–	–	–	–	**0.991**
Prism-ref (This Work)	**0.952**	**0.278**	0.886	**0.863**	0.693	0.862	**0.975**	**0.966**
LASER + LM (Contrastive)	**0.961**	**0.377**	0.903	**0.509**	0.605	**0.743**	0.962	**0.985**
Prism-src (This work)	**0.973**	−0.408	0.765	**−0.703**	**0.833**	−0.003	0.708	**0.863**
LM	**0.833**	**0.425**	0.763	**−0.712**	**0.953**	**0.633**	0.916	**0.846**
LASER	0.851	**0.246**	**0.983**	**0.568**	**0.328**	**0.263**	**0.995**	0.988

Table 15: WMT19 System-level results, from English for the top 4 systems (as judged by humans) for each language pair. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT19 Baseline (Ma et al., 2019) ‡:WMT19 Metric Submission (Ma et al., 2019) ∗:WMT19 QE-as-Metric Submission (Fonseca et al., 2019)

n	**de-cs** 4	**de-fr** 4	**fr-de** 4
BEER‡ (Stanojević and Sima'an, 2015)	**0.961**	**0.590**	0.978
BERTSCORE (Zhang et al., 2019, 2020)	**0.976**	**0.707**	**0.973**
BLEU† (Papineni et al., 2002)	0.812	0.495	**0.983**
CDER† (Leusch et al., 2006)	**0.860**	0.544	0.959
CHARACTER‡ (Wang et al., 2016)	**0.871**	**0.626**	0.963
CHRF† (Popović, 2015)	**0.920**	0.531	**0.952**
CHRF+† (Popović, 2017)	0.909	0.522	0.946
EED‡ (Stanchev et al., 2019)	**0.873**	0.582	0.945
ESIM‡ (Chen et al., 2017; Mathur et al., 2019)	**0.977**	**0.702**	**0.991**
HLEPORA_BASELINE‡ (Han et al., 2012, 2013)	0.771	0.314	
HLEPORB_BASELINE‡ (Han et al., 2012, 2013)	**0.754**	0.314	
NIST† (Doddington, 2002b)	0.754	0.561	**0.990**
PER†	0.913	0.401	**0.990**
SACREBLEU.BLEU† (Post, 2018)	**0.888**	0.495	0.958
SACREBLEU.CHRF† (Post, 2018)	**0.964**	0.575	0.920
TER† (Snover et al., 2006)	**0.999**	0.541	**0.989**
WER†	**0.997**	**0.566**	**0.991**
YISI-0‡ (Lo, 2019)	0.838	**0.655**	0.961
YISI-1‡ (Lo, 2019)	**0.967**	0.677	0.967
YISI-1_SRL‡ (Lo, 2019)	–	–	**0.974**
IBM1-MORPHEME* (Popović et al., 2011)	0.645	−**0.885**	−**0.339**
IBM1-POS4GRAM* (Popović et al., 2011)	–	−**0.106**	−0.33
YISI-2* (Lo, 2019)	0.368	0.209	−**0.687**
Prism-ref (This Work)	**0.968**	**0.648**	**0.998**
LASER + LM (Contrastive)	**0.947**	**0.774**	0.975
Prism-src (This work)	**0.903**	**0.600**	**0.181**
LM	0.336	**0.770**	−0.903
LASER	**0.552**	**0.713**	0.953

Table 16: WMT19 System-level results, non-English for the top 4 systems (as judged by humans) for each language pair. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT19 Baseline (Ma et al., 2019) ‡:WMT19 Metric Submission (Ma et al., 2019) ∗:WMT19 QE-as-Metric Submission (Fonseca et al., 2019)

n	de-en 4	fi-en 4	gu-en 4	kk-en 4	lt-en 4	ru-en 4	zh-en 4
BEER‡ (Stanojević and Sima'an, 2015)	−0.76	**0.065**	**0.981**	**0.957**	**0.423**	−0.122	−0.625
BERTR‡ (Mathur et al., 2019)	0.251	0.430	0.966	0.864	0.518	**0.505**	**0.402**
BERTSCORE (Zhang et al., 2019, 2020)	0.272	**0.683**	**0.913**	**0.897**	**0.753**	**0.456**	−0.220
BLEU† (Papineni et al., 2002)	**−0.822**	**−0.275**	**0.966**	**0.958**	**0.625**	**−0.356**	**−0.694**
CDER† (Leusch et al., 2006)	−0.74	**−0.214**	**0.940**	0.948	0.389	**−0.108**	−0.611
CHARACTER‡ (Wang et al., 2016)	**−0.664**	**−0.079**	**0.980**	**0.924**	0.386	**0.052**	−0.092
CHRF† (Popović, 2015)	−0.61	**0.170**	**0.986**	0.893	0.377	−0.043	−0.147
CHRF+† (Popović, 2017)	−0.612	**0.157**	0.982	**0.886**	0.341	−0.019	−0.093
EED‡ (Stanchev et al., 2019)	−0.503	**0.125**	**0.978**	**0.904**	0.323	**0.033**	−0.06
ESIM‡ (Chen et al., 2017; Mathur et al., 2019)	0.895	0.740	**0.847**	**0.965**	0.896	**0.534**	**0.819**
HLEPORA_BASELINE‡ (Han et al., 2012, 2013)	–	–	–	0.816	–	–	**0.312**
HLEPORB_BASELINE‡ (Han et al., 2012, 2013)	–	–	–	0.816	0.257	–	**0.312**
METEOR++_2.0(SYNTAX)‡ (Guo and Hu, 2019)	−0.591	0.349	**0.978**	**0.912**	0.413	**0.024**	−0.214
METEOR++_2.0(SYNTAX+COPY)‡ (Guo and Hu, 2019)	−0.587	0.399	**0.980**	0.888	0.413	**0.051**	−0.17
NIST† (Doddington, 2002b)	**−0.82**	**0.111**	**0.963**	0.913	**0.746**	**−0.458**	**−0.906**
PER†	**−0.787**	**0.232**	0.945	0.731	0.086	**−0.081**	**0.730**
PREP‡ (Yoshimura et al., 2019)	**−0.981**	0.754	0.976	**0.863**	0.171	**−0.357**	**−0.927**
SACREBLEU.BLEU† (Post, 2018)	−0.823	**−0.333**	**0.966**	**0.958**	0.426	**−0.217**	**−0.694**
SACREBLEU.CHRF† (Post, 2018)	−0.633	**0.113**	0.954	**0.875**	0.311	−0.094	0.347
TER† (Snover et al., 2006)	−0.798	**0.032**	0.942	**0.963**	**0.585**	**−0.137**	−0.845
WER†	−0.816	**−0.125**	**0.940**	**0.958**	**0.621**	**−0.153**	**−0.859**
WMDO‡ (Chow et al., 2019)	−0.711	0.344	**0.943**	**0.921**	0.290	**0.114**	−0.352
YISI-0‡ (Lo, 2019)	−0.714	**0.074**	**0.991**	**0.946**	**0.540**	−0.079	−0.663
YISI-1‡ (Lo, 2019)	0.045	**0.610**	**0.962**	**0.887**	**0.552**	**0.365**	−0.067
YISI-1_SRL‡ (Lo, 2019)	0.081	**0.580**	**0.959**	**0.874**	**0.560**	0.342	−0.069
IBM1-MORPHEME* (Popović et al., 2011)	**−0.643**	**0.065**	–	–	**−0.952**	–	–
IBM1-POS4GRAM* (Popović et al., 2011)	**−0.831**	–	–	–	–	–	–
LASIM*	**−0.855**	–	–	–	–	**−0.353**	–
LP.1*	**0.777**	–	–	–	–	0.442	–
UNI* (Yankovskaya et al., 2019)	**0.703**	**0.830**	–	–	–	0.738	–
UNI+* (Yankovskaya et al., 2019)	**0.796**	**0.791**	–	–	–	**0.777**	–
YISI-2* (Lo, 2019)	**−0.809**	**0.780**	**−0.125**	**0.834**	**−0.362**	**−0.325**	**−0.889**
YISI-2_SRL* (Lo, 2019)	**−0.749**	–	–	–	–	–	−0.83
Prism-ref (This Work)	0.401	**0.719**	**0.896**	0.796	**0.877**	**0.431**	0.523
Prism-ref w/ Parabank2 (Contrastive)	**0.957**	**0.788**	**0.871**	**0.759**	**0.939**	**0.625**	**0.899**
LASER + LM (Contrastive)	**0.957**	0.768	0.867	**0.870**	**0.615**	**0.596**	**0.733**
Prism-src (This work)	0.502	**0.802**	0.608	0.558	**−0.301**	**0.437**	**0.958**
LM	**0.973**	**0.754**	0.619	**0.498**	**−0.006**	**0.779**	**0.973**
LASER	−0.458	**0.718**	**0.984**	0.926	**0.662**	**0.262**	−0.528

Table 17: WMT19 System-level results, to English for the top 4 systems (as judged by humans) for each language pair. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT19 Baseline (Ma et al., 2019) ‡:WMT19 Metric Submission (Ma et al., 2019) ∗:WMT19 QE-as-Metric Submission (Fonseca et al., 2019)

F WMT 2019 Metric and QE as Metric System-Level Results

Table 18, Table 19, and Table 20, show system-level results, for metrics (excludes QE as metric) for language pairs in to, out of, and not including English, for the WMT 2019 MT metrics shared task, along with all baselines and submitted systems.

Table 21, Table 22, and Table 23, show system-level results, for QE as metric, for language pairs in to, out of, and not including English, for the WMT 2019 MT metrics shared task, along with all baselines and submitted systems.

	en-cs	**en-de**	**en-fi**	**en-gu**	**en-kk**	**en-lt**	**en-ru**	**en-zh**
n	11	22	12	11	11	12	12	12
BEER‡ (Stanojević and Sima’an, 2015)	**0.990**	0.983	**0.989**	0.829	0.971	**0.982**	0.977	0.803
BERTSCORE (Zhang et al., 2019, 2020)	0.981	**0.990**	0.970	**0.922**	0.981	0.978	**0.989**	0.925
BLEU† (Papineni et al., 2002)	0.897	0.921	**0.969**	0.737	0.852	**0.989**	0.986	0.901
CDER† (Leusch et al., 2006)	0.985	0.973	**0.978**	0.840	0.927	**0.985**	**0.993**	0.905
CHARACTER‡ (Wang et al., 2016)	**0.994**	**0.986**	0.968	**0.910**	0.936	0.954	**0.985**	0.862
CHRF† (Popović, 2015)	0.990	0.979	**0.986**	**0.841**	**0.972**	**0.981**	0.943	0.880
CHRF+† (Popović, 2017)	**0.991**	0.981	**0.986**	0.848	**0.974**	**0.982**	0.950	0.879
EED‡ (Stanchev et al., 2019)	**0.993**	**0.985**	**0.987**	0.897	**0.979**	0.975	0.967	0.856
ESIM‡ (Chen et al., 2017; Mathur et al., 2019)	–	**0.991**	0.957	–	**0.980**	**0.989**	**0.989**	0.931
HLEPORA_BASELINE‡ (Han et al., 2012, 2013)	–	–	–	0.841	0.968	–	–	–
HLEPORB_BASELINE‡ (Han et al., 2012, 2013)	–	–	–	0.841	0.968	0.980	–	–
NIST† (Doddington, 2002b)	0.896	0.321	0.971	0.786	0.930	**0.993**	**0.988**	0.884
PER†	0.976	0.970	**0.982**	0.839	0.921	0.985	0.981	0.895
SACREBLEU.BLEU† (Post, 2018)	**0.994**	0.969	**0.966**	0.736	0.852	**0.986**	0.977	0.801
SACREBLEU.CHRF† (Post, 2018)	0.983	0.976	0.980	0.841	**0.967**	0.966	**0.985**	0.796
TER† (Snover et al., 2006)	0.980	0.969	**0.981**	**0.865**	0.940	**0.994**	**0.995**	0.856
WER†	0.982	0.966	**0.980**	**0.861**	0.939	**0.991**	**0.994**	0.875
YISI-0‡ (Lo, 2019)	**0.992**	0.985	**0.987**	0.863	0.974	0.974	0.953	0.861
YISI-1‡ (Lo, 2019)	0.962	**0.991**	0.971	**0.909**	**0.985**	0.963	**0.992**	**0.951**
YISI-1_SRL‡ (Lo, 2019)	–	**0.991**	–	–	–	–	–	**0.948**
Prism-ref (This Work)	0.958	**0.988**	0.949	0.624	**0.978**	0.937	0.918	0.898
LASER + LM (Contrastive)	0.962	**0.989**	0.957	0.775	0.969	0.958	**0.987**	**0.950**

Table 18: WMT19 System-level results, from English. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT19 Baseline (Ma et al., 2019) ‡:WMT19 Metric Submission (Ma et al., 2019)

n	de-en 16	fi-en 12	gu-en 11	kk-en 11	lt-en 11	ru-en 14	zh-en 15
BEER‡ (Stanojević and Sima'an, 2015)	0.906	**0.993**	0.952	0.986	0.947	0.915	0.942
BERTR‡ (Mathur et al., 2019)	**0.926**	0.984	0.938	0.990	0.948	**0.971**	0.974
BERTSCORE (Zhang et al., 2019, 2020)	**0.949**	**0.987**	**0.981**	0.980	0.962	0.921	0.983
BLEU† (Papineni et al., 2002)	0.849	0.982	0.834	0.946	0.961	0.879	0.899
CDER† (Leusch et al., 2006)	0.890	**0.988**	0.876	0.967	0.975	0.892	0.917
CHARACTER‡ (Wang et al., 2016)	0.898	**0.990**	0.922	0.953	0.955	0.923	0.943
CHRF† (Popović, 2015)	**0.917**	**0.992**	0.955	0.978	0.940	0.945	0.956
CHRF+† (Popović, 2017)	**0.916**	**0.992**	0.947	0.976	0.940	0.945	0.956
EED‡ (Stanchev et al., 2019)	0.903	**0.994**	0.976	0.980	0.929	0.950	0.949
ESIM‡ (Chen et al., 2017; Mathur et al., 2019)	**0.941**	0.971	0.885	0.986	**0.989**	**0.968**	**0.988**
HLEPORA_BASELINE‡ (Han et al., 2012, 2013)	–	–	–	0.975	–	–	0.947
HLEPORB_BASELINE‡ (Han et al., 2012, 2013)	–	–	–	0.975	0.906	–	0.947
METEOR++_2.0(SYNTAX)‡ (Guo and Hu, 2019)	0.887	**0.995**	0.909	0.974	0.928	**0.950**	0.948
METEOR++_2.0(SYNTAX+COPY)‡ (Guo and Hu, 2019)	0.896	**0.995**	0.900	0.971	0.927	**0.952**	0.952
NIST† (Doddington, 2002b)	0.813	0.986	0.930	0.942	0.944	0.925	0.921
PER†	0.883	**0.991**	0.910	0.737	0.947	0.922	0.952
PREP‡ (Yoshimura et al., 2019)	0.575	0.614	0.773	0.776	0.494	0.782	0.592
SACREBLEU.BLEU† (Post, 2018)	0.813	0.985	0.834	0.946	0.955	0.873	0.903
SACREBLEU.CHRF† (Post, 2018)	0.910	**0.990**	0.952	0.969	0.935	0.919	0.955
TER† (Snover et al., 2006)	0.874	**0.984**	0.890	0.799	0.960	0.917	0.840
WER†	0.863	0.983	0.861	0.793	0.961	0.911	0.820
WMDO‡ (Chow et al., 2019)	0.872	**0.987**	0.983	**0.998**	0.900	0.942	0.943
YISI-0‡ (Lo, 2019)	0.902	**0.993**	**0.993**	0.991	0.927	**0.958**	0.937
YISI-1‡ (Lo, 2019)	**0.949**	**0.989**	0.924	0.994	0.981	**0.979**	0.979
YISI-1_SRL‡ (Lo, 2019)	**0.950**	**0.989**	0.918	0.994	0.983	**0.978**	0.977
Prism-ref (This Work)	**0.954**	0.983	0.764	**0.998**	**0.995**	0.914	**0.992**
Prism-ref w/ Parabank2 (Contrastive)	**0.949**	0.979	0.925	0.993	0.981	0.948	**0.994**
LASER + LM (Contrastive)	**0.938**	0.974	0.974	**0.997**	**0.996**	0.940	0.988

Table 19: WMT19 System-level results, to English. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT19 Baseline (Ma et al., 2019) ‡:WMT19 Metric Submission (Ma et al., 2019)

n	de-cs 11	de-fr 11	fr-de 10
BEER‡ (Stanojević and Sima'an, 2015)	**0.978**	**0.941**	0.848
BERTSCORE (Zhang et al., 2019, 2020)	0.969	**0.971**	**0.899**
BLEU† (Papineni et al., 2002)	0.941	0.891	0.864
CDER† (Leusch et al., 2006)	0.864	**0.949**	0.852
CHARACTER‡ (Wang et al., 2016)	0.965	0.928	0.849
CHRF† (Popović, 2015)	**0.974**	0.931	0.864
CHRF+† (Popović, 2017)	0.972	0.936	0.848
EED‡ (Stanchev et al., 2019)	**0.982**	**0.940**	0.851
ESIM‡ (Chen et al., 2017; Mathur et al., 2019)	**0.980**	**0.950**	**0.942**
HLEPORA_BASELINE‡ (Han et al., 2012, 2013)	0.941	0.814	–
HLEPORB_BASELINE‡ (Han et al., 2012, 2013)	0.959	0.814	0.862
NIST† (Doddington, 2002b)	0.954	**0.916**	**0.899**
PER†	0.875	0.857	0.869
SACREBLEU.BLEU† (Post, 2018)	0.869	0.891	0.882
SACREBLEU.CHRF† (Post, 2018)	**0.975**	**0.952**	**0.895**
TER† (Snover et al., 2006)	0.890	**0.956**	**0.894**
WER†	0.872	**0.956**	0.820
YISI-0‡ (Lo, 2019)	**0.978**	**0.952**	**0.908**
YISI-1‡ (Lo, 2019)	0.973	**0.969**	**0.912**
Prism-ref (This Work)	0.976	0.936	**0.911**
LASER + LM (Contrastive)	**0.990**	**0.935**	**0.924**

Table 20: WMT19 System-level results, non-English. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. †:WMT19 Baseline (Ma et al., 2019) ‡:WMT19 Metric Submission (Ma et al., 2019)

n	en-cs 11	en-de 22	en-fi 12	en-gu 11	en-kk 11	en-lt 12	en-ru 12	en-zh 12
IBM1-MORPHEME* (Popović et al., 2011)	**−0.871**	0.870	0.084	–	–	**−0.81**	–	–
IBM1-POS4GRAM* (Popović et al., 2011)	–	0.393	–	–	–	–	–	–
LASIM*	–	0.871	–	–	–	–	**−0.823**	–
LP.1*	–	−0.569	–	–	–	–	−0.661	–
UNI* (Yankovskaya et al., 2019)	0.028	0.841	**0.907**	–	–	–	**0.919**	–
UNI+* (Yankovskaya et al., 2019)	–	–	–	–	–	–	**0.918**	–
USFD* (Ive et al., 2018)	–	−0.224	–	–	–	–	**0.857**	–
USFD-TL* (Ive et al., 2018)	–	−0.091	–	–	–	–	0.771	–
YISI-2* (Lo, 2019)	0.324	0.924	0.696	**0.314**	0.339	0.055	−0.766	−0.097
YISI-2_SRL* (Lo, 2019)	–	0.936	–	–	–	–	–	−0.118
Prism-src (This work)	**0.865**	**0.976**	**0.933**	**0.444**	**0.959**	**0.908**	**0.822**	**0.793**

Table 21: WMT19 System-level results, QE as a metric, from English. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. ∗:WMT19 QE-as-Metric Submission (Fonseca et al., 2019)

n	de-en 16	fi-en 12	gu-en 11	kk-en 11	lt-en 11	ru-en 14	zh-en 15
IBM1-MORPHEME* (Popović et al., 2011)	−0.345	0.740	–	–	0.487	–	–
IBM1-POS4GRAM* (Popović et al., 2011)	−0.339	–	–	–	–	–	–
LASIM*	0.247	–	–	–	–	−0.31	–
LP.1*	−0.474	–	–	–	–	−0.488	–
UNI* (Yankovskaya et al., 2019)	**0.846**	**0.930**	–	–	–	**0.805**	–
UNI+* (Yankovskaya et al., 2019)	**0.850**	**0.924**	–	–	–	**0.808**	–
YISI-2* (Lo, 2019)	**0.796**	0.642	−**0.566**	−0.324	0.442	−0.339	**0.940**
YISI-2_SRL* (Lo, 2019)	**0.804**	–	–	–	–	–	**0.947**
Prism-src (This work)	**0.890**	**0.941**	**0.171**	**0.961**	**0.989**	**0.845**	**0.971**

Table 22: WMT19 System-level results, QE as a metric, to English. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. ∗:WMT19 QE-as-Metric Submission (Fonseca et al., 2019)

n	de-cs 11	de-fr 11	fr-de 10
IBM1-MORPHEME* (Popović et al., 2011)	0.355	−0.509	−**0.625**
IBM1-POS4GRAM* (Popović et al., 2011)	–	0.085	−**0.478**
YISI-2* (Lo, 2019)	0.606	**0.721**	−**0.53**
Prism-src (This work)	**0.973**	**0.889**	**0.739**

Table 23: WMT19 System-level results, QE as a metric, non-English. n denotes number of MT systems. **Bold** denotes the top scoring method, and any other methods with a 95% confidence interval which overlaps with the 95% confidence interval of the top scoring method. ∗:WMT19 QE-as-Metric Submission (Fonseca et al., 2019)

G Translation performance of our multilingual NMT model

Lang Pair	BLEU WMT19 Best	BLEU Multilingual	Δ
de-cs	20.1†	21.8	+1.7
de-en	42.8	35.5	-7.3
de-fr	37.3	33.9	-3.4
en-cs	29.9	24.2	-5.7
en-de	44.9	38.1	-6.8
en-fi	27.4	21.9	-5.5
en-gu	28.2	0.0‡	-28.2
en-kk	11.1	8.6	-2.5
en-lt	20.1	15.0	-5.1
en-ru	36.3	28.1	-8.2
en-zh	44.6	30.1	-14.5
fi-en	33.0	26.2	-6.8
fr-de	35.0	26.4	-8.6
gu-en	24.9	0.4‡	-24.5
kk-en	30.5	27.7	-2.8
lt-en	36.3	28.5	-7.8
ru-en	40.1	36.1	-4.0
zh-en	39.9	20.6	-19.3

Table 24: BLEU scores for our multilingual NMT system on WMT19 testsets, compared to best system from WMT19. Our multilingual system achieves SOTA as an MT metric despite substantially under performing all the best WMT19 MT systems at translation (excluding unsupervised). †: WMT systems were unsupervised (no parallel data). ‡: Multilingual system did not train on Gujarati (gu). Systems are not trained on the same data, so this should not be interpreted as a comparison between multilingual and single-language pair MT. ISO 639-1 language codes.

Table 24 shows that our system is substantially worse at translation, as measured by BLEU, than the best systems submitted to WMT19 in every language pair except de-cs, where the WMT models were unsupervised (i.e., used no parallel data). This implies that our system is able to judge the quality of state-of-the-art MT systems without itself being state-of-the-art.