On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation

Wei Zhao[†], **Goran Glavaš**[‡], **Maxime Peyrard**[Φ], **Yang Gao**[⋆], **Robert West**[Φ], **Steffen Eger**[†]
[†] Technische Universität Darmstadt [‡] University of Mannheim, Germany
[Φ] EPFL, Switerland [⋆] Royal Holloway University of London, UK
{zhao,eger}@aiphes.tu-darmstadt.de
goran@informatik.uni-mannheim.de, yang.gao@rhul.ac.uk
{maxime.peyrard,robert.west}@epfl.ch

Abstract

Evaluation of cross-lingual encoders is usually performed either via zero-shot cross-lingual transfer in supervised downstream tasks or via unsupervised cross-lingual textual similarity. In this paper, we concern ourselves with reference-free machine translation (MT) evaluation where we directly compare source texts to (sometimes low-quality) system translations, which represents a natural adversarial setup for multilingual encoders. Reference-free evaluation holds the promise of web-scale comparison of MT systems. We systematically investigate a range of metrics based on state-of-the-art cross-lingual semantic representations obtained with pretrained M-BERT and LASER. We find that they perform poorly as semantic encoders for reference-free MT evaluation and identify their two key limitations, namely, (a) a semantic mismatch between representations of mutual translations and, more prominently, (b) the inability to punish "translationese", i.e., low-quality literal translations. We propose two partial remedies: (1) post-hoc re-alignment of the vector spaces and (2) coupling of semantic-similarity based metrics with target-side language modeling. In segment-level MT evaluation, our best metric surpasses reference-based BLEU by 5.7 correlation points. We make our MT evaluation code available.[1]

1 Introduction

A standard evaluation setup for supervised machine learning (ML) tasks assumes an evaluation metric which compares a gold label to a classifier prediction. This setup assumes that the task has clearly defined and unambiguous labels and, in most cases, that an instance can be assigned few labels. These assumptions, however, do not hold for natural language generation (NLG) tasks like machine translation (MT) (Bahdanau et al., 2015; Johnson et al., 2017) and text summarization (Rush et al., 2015; Tan et al., 2017), where we do not predict a single discrete label but generate natural language text. Thus, the set of labels for NLG is neither clearly defined nor finite. Yet, the standard evaluation protocols for NLG still predominantly follow the described default paradigm: (1) evaluation datasets come with human-created reference texts and (2) evaluation metrics, e.g., BLEU (Papineni et al., 2002) or METEOR (Lavie and Agarwal, 2007) for MT and ROUGE (Lin and Hovy, 2003) for summarization, count the exact "label" (i.e., n-gram) matches between reference and system-generated text. In other words, established NLG evaluation compares semantically ambiguous labels from an unbounded set (i.e., natural language texts) via hard symbolic matching (i.e., string overlap).

The first remedy is to replace the hard symbolic comparison of natural language "labels" with a soft comparison of texts' meaning, using semantic vector space representations. Recently, a number of MT evaluation methods appeared focusing on semantic comparison of reference and system translations (Shimanaka et al., 2018; Clark et al., 2019; Zhao et al., 2019). While these correlate better than n-gram overlap metrics with human assessments, they do not address inherent limitations stemming from the need for reference translations, namely: (1) references are expensive to obtain; (2) they assume a single correct solution and bias the evaluation, both automatic and human (Dreyer and Marcu, 2012; Fomicheva and Specia, 2016), and (3) limitation of MT evaluation to language pairs with available parallel data.

Reliable *reference-free* evaluation metrics, directly measuring the (semantic) correspondence between the source language text and system translation, would remove the need for human references and allow for unlimited MT evaluations: *any*

[1] https://github.com/AIPHES/ACL20-Reference-Free-MT-Evaluation

monolingual corpus could be used for evaluating MT systems. However, the proposals of reference-free MT evaluation metrics have been few and far apart and have required either non-negligible supervision (i.e., human translation quality labels) (Specia et al., 2010) or language-specific preprocessing like semantic parsing (Lo et al., 2014; Lo, 2019), both hindering the wide applicability of the proposed metrics. Moreover, they have also typically exhibited performance levels well below those of standard reference-based metrics (Ma et al., 2019).

In this work, we comparatively evaluate a number of reference-free MT evaluation metrics that build on the most recent developments in multilingual representation learning, namely cross-lingual contextualized embeddings (Devlin et al., 2019) and cross-lingual sentence encoders (Artetxe and Schwenk, 2019). We investigate two types of cross-lingual reference-free metrics: **(1)** *Soft token-level alignment* metrics find the optimal soft alignment between source sentence and system translation using Word Mover's Distance (WMD) (Kusner et al., 2015). Zhao et al. (2019) recently demonstrated that WMD operating on BERT representations (Devlin et al., 2019) substantially outperforms baseline MT evaluation metrics in the reference-based setting. In this work, we investigate whether WMD can yield comparable success in the reference-free (i.e., cross-lingual) setup; **(2)** *Sentence-level similarity* metrics measure the similarity between sentence representations of the source sentence and system translation using cosine similarity.

Our analysis yields several interesting findings. **(i)** We show that, unlike in the monolingual reference-based setup, metrics that operate on contextualized representations generally do not outperform symbolic matching metrics like BLEU, which operate in the reference-based environment. **(ii)** We identify two reasons for this failure: (a) firstly, cross-lingual semantic mismatch, especially for multi-lingual BERT (M-BERT), which construes a shared multilingual space in an unsupervised fashion, without any direct bilingual signal; (b) secondly, the inability of the state-of-the-art cross-lingual metrics based on multilingual encoders to adequately capture and punish "translationese", i.e., literal word-by-word translations of the source sentence—as translationese is an especially persistent property of MT systems, this problem is particularly troubling in our context of reference-free MT evaluation. **(iii)** We show that by executing an additional weakly-supervised cross-lingual re-mapping step, we can to some extent alleviate both previous issues. **(iv)** Finally, we show that the combination of cross-lingual reference-free metrics and language modeling on the target side (which is able to detect "translationese"), surpasses the performance of reference-based baselines.

Beyond designating a viable prospect of web-scale domain-agnostic MT evaluation, our findings indicate that the challenging task of reference-free MT evaluation is able to expose an important limitation of current state-of-the-art multilingual encoders, i.e., the failure to properly represent corrupt input, that may go unnoticed in simpler evaluation setups such as zero-shot cross-lingual text classification or measuring cross-lingual text similarity not involving "adversarial" conditions. We believe this is a promising direction for nuanced, fine-grained evaluation of cross-lingual representations, extending the recent benchmarks which focus on zero-shot transfer scenarios (Hu et al., 2020).

2 Related Work

Manual human evaluations of MT systems undoubtedly yield the most reliable results, but are expensive, tedious, and generally do not scale to a multitude of domains. A significant body of research is thus dedicated to the study of automatic evaluation metrics for machine translation. Here, we provide an overview of both reference-based MT evaluation metrics and recent research efforts towards reference-free MT evaluation, which leverage cross-lingual semantic representations and unsupervised MT techniques.

Reference-based MT evaluation. Most of the commonly used evaluation metrics in MT compare system and reference translations. They are often based on surface forms such as n-gram overlaps like BLEU (Papineni et al., 2002), SentBLEU, NIST (Doddington, 2002), chrF++ (Popović, 2017) or METEOR++(Guo and Hu, 2019). They have been extensively tested and compared in recent WMT metrics shared tasks (Bojar et al., 2017a; Ma et al., 2018a, 2019).

These metrics, however, operate at the surface level, and by design fail to recognize semantic equivalence lacking lexical overlap. To overcome these limitations, some research efforts exploited static word embeddings (Mikolov et al., 2013b) and trained embedding-based supervised metrics on sufficiently large datasets with available human judgments of translation quality (Shimanaka

et al., 2018). With the development of contextual word embeddings (Peters et al., 2018; Devlin et al., 2019), we have witnessed proposals of semantic metrics that account for word order. For example, Clark et al. (2019) introduce a semantic metric relying on sentence movers similarity and the contextualized ELMo embeddings (Peters et al., 2018). Similarly, Zhang et al. (2019) describe a reference-based semantic similarity metric based on contextualized BERT representations (Devlin et al., 2019). Zhao et al. (2019) generalize this line of work with their MoverScore metric, which computes the mover's distance, i.e., the optimal soft alignment between tokens of the two sentences, based on the similarities between their contextualized embeddings. Mathur et al. (2019) train a supervised BERT-based regressor for reference-based MT evaluation.

Reference-free MT evaluation. Recently, there has been a growing interest in reference-free MT evaluation (Ma et al., 2019), also referred to as "quality estimation" (QE) in the MT community. In this setup, evaluation metrics semantically compare system translations directly to the source sentences. The attractiveness of automatic reference-free MT evaluation is obvious: it does not require any human effort or parallel data. To approach this task, Popović et al. (2011) exploit a bag-of-word translation model to estimate translation quality, which sums over the likelihoods of aligned word-pairs between source and translation texts. Specia et al. (2013) estimate translation quality using language-agnostic linguistic features extracted from source lanuage texts and system translations. Lo et al. (2014) introduce XMEANT as a cross-lingual reference-free variant of MEANT, a metric based on semantic frames. Lo (2019) extended this idea by leveraging M-BERT embeddings. The resulting metric, YiSi-2, evaluates system translations by summing similarity scores over words pairs that are best-aligned mutual translations. YiSi-2-SRL optionally combines an additional similarity score based on the alignment over the semantic structures (e.g., semantic roles and frames). Both metrics are reference-free, but YiSi-2-SRL is not resource-lean as it requires a semantic parser for both languages. Moreover, in contrast to our proposed metrics, they do not mitigate the misalignment of cross-lingual embedding spaces and do not integrate a target-side language model, which we identify to be crucial components.

Recent progress in cross-lingual semantic similarity (Agirre et al., 2016; Cer et al., 2017) and unsupervised MT (Artetxe and Schwenk, 2019) has also led to novel reference-free metrics. For instance, Yankovskaya et al. (2019) propose to train a metric combining multilingual embeddings extracted from M-BERT and LASER (Artetxe and Schwenk, 2019) together with the log-probability scores from neural machine translation. Our work differs from that of Yankovskaya et al. (2019) in one crucial aspect: the cross-lingual reference-free metrics that we investigate and benchmark do not require any human supervision.

Cross-lingual Representations. Cross-lingual text representations offer a prospect of modeling meaning across languages and support cross-lingual transfer for downstream tasks (Klementiev et al., 2012; Rücklé et al., 2018; Glavaš et al., 2019; Josifoski et al., 2019; Conneau et al., 2020). Most recently, the (massively) multilingual encoders, such as multilingual M-BERT (Devlin et al., 2019), XLM-on-RoBERTa (Conneau et al., 2020), and (sentence-based) LASER, have profiled themselves as state-of-the-art solutions for (massively) multilingual semantic encoding of text. While LASER has been jointly trained on parallel data of 93 languages, M-BERT has been trained on the concatenation of monolingual data in more than 100 languages, without any cross-lingual mapping signal. There has been a recent vivid discussion on the cross-lingual abilities of M-BERT (Pires et al., 2019; K et al., 2020; Cao et al., 2020). In particular, Cao et al. (2020) show that M-BERT often yields disparate vector space representations for mutual translations and propose a multilingual re-mapping based on parallel corpora, to remedy for this issue. In this work, we introduce re-mapping solutions that are resource-leaner and require easy-to-obtain limited-size word translation dictionaries rather than large parallel corpora.

3 Reference-Free MT Evaluation Metrics

In the following, we use $\mathbf{x}$ to denote a source sentence (i.e., a sequence of tokens in the source language), $\mathbf{y}$ to denote a system translation of $\mathbf{x}$ in the target language, and $\mathbf{y}^{\star}$ to denote the human reference translation for $\mathbf{x}$.

3.1 Soft Token-Level Alignment

We start from the MoverScore (Zhao et al., 2019), a recently proposed reference-based MT evaluation

metric designed to measure the semantic similarity between system outputs ($\mathbf{y}$) and human references ($\mathbf{y}^\star$). It finds an optimal soft semantic alignments between tokens from $\mathbf{y}$ and $\mathbf{y}^\star$ by minimizing the Word Mover's Distance (Kusner et al., 2015). In this work, we extend the MoverScore metric to operate in the cross-lingual setup, i.e., to measure the semantic similarity between n-grams (unigram or bigrams) of the source text $\mathbf{x}$ and the system translation $\mathbf{y}$, represented with embeddings originating from a cross-lingual semantic space.

First, we decompose the source text $\mathbf{x}$ into a sequence of n-grams, denoted by $\boldsymbol{x}_n = (\mathrm{x}_1^n, \ldots, \mathrm{x}_m^n)$ and then do the same operation for the system translation $\mathbf{y}$, denoting the resulting sequence of n-grams with $\boldsymbol{y}_n$. Given $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$, we can then define a distance matrix $\boldsymbol{C}$ such that $\boldsymbol{C}_{ij} = \|E(\mathrm{x}_i^n) - E(\mathrm{y}_j^n)\|_2$ is the distance between the i-th n-gram of $\boldsymbol{x}$ and the j-th n-gram of $\boldsymbol{y}$, where E is a *cross-lingual* embedding function that maps text in different languages to a shared embedding space. With respect to the function E, we experimented with cross-lingual representations induced (a) from static word embeddings with RCSLS (Joulin et al., 2018)) (b) with M-BERT (Devlin et al., 2019) as the multilingual encoder; with a focus on the latter. For M-BERT, we take the representations of the last transformer layer as the text representations.

WMD between the two sequences of n-grams $\boldsymbol{x}^n$ and $\boldsymbol{y}^n$ with associated n-gram weights [2] to $\boldsymbol{f}_{\boldsymbol{x}^n} \in \mathbb{R}^{|\boldsymbol{x}^n|}$ and $\boldsymbol{f}_{\boldsymbol{y}^n} \in \mathbb{R}^{|\boldsymbol{y}^n|}$ is defined as:

$$m(\boldsymbol{x}, \boldsymbol{y}) := \mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) = \min_{\boldsymbol{F}} \sum_{ij} \boldsymbol{C}_{ij} \cdot \boldsymbol{F}_{ij},$$

$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^\intercal \mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n},$$

where $\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}$ is a transportation matrix with $\boldsymbol{F}_{ij}$ denoting the amount of flow traveling from x_i^n to y_j^n.

3.2 Sentence-Level Semantic Similarity

In addition to measuring semantic distance between $\mathbf{x}$ and $\mathbf{y}$ at word-level, one can also encode them into sentence representations with multilingual sentence encoders like LASER (Artetxe and Schwenk, 2019), and then measure their cosine distance

$$m(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{E(\boldsymbol{x})^\intercal E(\boldsymbol{y})}{\|E(\boldsymbol{x})\| \cdot \|E(\boldsymbol{y})\|}.$$

[2] We follow Zhao et al. (2019) in obtaining n-gram embeddings and their associated weights based on IDF.

3.3 Improving Cross-Lingual Alignments

Initial analysis indicated that, despite the multilingual pretraining of M-BERT (Devlin et al., 2019) and LASER (Artetxe and Schwenk, 2019), the monolingual subspaces of the multilingual spaces they induce are far from being semantically well-aligned, i.e., we obtain fairly distant vectors for mutual word or sentence translations.[3] To this end, we apply two simple, weakly-supervised linear projection methods for post-hoc improvement of the cross-lingual alignments in these multilingual representation spaces.

Notation. Let $\boldsymbol{D} = \{(w_\ell^1, w_k^1), \ldots, (w_\ell^n, w_k^n)\}$ be a set of matched word or sentence pairs from two different languages ℓ and k. We define a re-mapping function f such that any $f(E(w_\ell))$ and $E(w_k)$ are better aligned in the resulting shared vector space. We investigate two resource-lean choices for the re-mapping function f.

Linear Cross-lingual Projection (CLP). Following related work (Schuster et al., 2019), we re-map contextualized embedding spaces using linear projection. Given ℓ and k, we stack all vectors of the source language words and target language words for pairs $\boldsymbol{D}$, respectively, to form matrices $\boldsymbol{X}_\ell$ and $\boldsymbol{X}_k \in \mathbb{R}^{n \times d}$, with d as the embedding dimension and n as the number of word or sentence alignments. The word pairs we use to calibrate M-BERT are extracted from EuroParl (Koehn, 2005) using FastAlign (Dyer et al., 2013), and the sentence pairs to calibrate LASER are sampled directly from EuroParl.[4] Mikolov et al. (2013a) propose to learn a projection matrix $W \in \mathbb{R}^{d \times d}$ by minimizing the Euclidean distance beetween the projected source language vectors and their corresponding target language vectors:

$$\min_{\boldsymbol{W}} \|\boldsymbol{W}\boldsymbol{X}_\ell - \boldsymbol{X}_k\|_2.$$

Xing et al. (2015) achieve further improvement on the task of bilingual lexicon induction (BLI) by constraining $\boldsymbol{W}$ to an orthogonal matrix, i.e., such that $\boldsymbol{W}^\intercal\boldsymbol{W} = \mathbf{I}$. This turns the optimization into the well-known Procrustes problem (Schönemann, 1966) with the following closed-form solution:

$$\hat{\boldsymbol{W}} = \boldsymbol{U}\boldsymbol{V}^\intercal, \boldsymbol{U}\Sigma\boldsymbol{V}^\intercal = \mathrm{SVD}(\boldsymbol{X}_\ell\boldsymbol{X}_k^\intercal)$$

[3] LASER is jointly trained on parallel corpora of different languages, but in resource-lean language pairs, the induced embeddings from mutual translations may be far apart.

[4] While LASER requires large parallel corpora in pretraining, we believe that fine-tuning/calibrating the embeddings post-hoc requires fewer data points.

We note that the above CLP re-mapping is known to have deficits, i.e., it requires the embedding spaces of the involved languages to be approximately isomorphic (Søgaard et al., 2018; Vulić et al., 2019). Recently, some re-mapping methods that reportedly remedy for this issue have been suggested (Glavaš and Vulić, 2020; Mohiuddin and Joty, 2020). We leave the investigation of these novel techniques for our future work.

Universal Language Mismatch-Direction (UMD) Our second post-hoc linear alignment method is inspired by the recent work on removing biases in distributional word vectors (Dev and Phillips, 2019; Lauscher et al., 2019). We adopt the same approaches in order to quantify and remedy for the "language bias", i.e., representation mismatches between mutual translations in the initial multilingual space. Formally, given ℓ and k, we create individual misalignment vectors $E(w_\ell^i) - E(w_k^i)$ for each bilingual pair in $\boldsymbol{D}$. Then we stack these individual vectors to form a matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times d}$. We then obtain the global misalignment vector $\boldsymbol{v}_B$ as the top left singular vector of $\boldsymbol{Q}$. The global misalignment vector presumably captures the direction of the representational misalignment between the languages better than the individual (noisy) misalignment vectors $E(w_\ell^i) - E(w_k^i)$. Finally, we modify all vectors $E(w_\ell)$ and $E(w_k)$, by subtracting their projections onto the global misalignment direction vector v_B:

$$f(E(w_\ell)) = E(w_\ell) - \cos(E(w_\ell), v_B)v_B.$$

Language Model BLEU scores often fail to reflect the fluency level of translated texts (Edunov et al., 2019). Hence, we use the language model (LM) of the target language to regularize the cross-lingual semantic similarity metrics, by coupling our cross-lingual similarity scores with a GPT language model of the target language (Radford et al., 2018). We expect the language model to penalize translationese, i.e., unnatural word-by-word translations and boost the performance of our metrics.[5]

4 Experiments

In this section, we evaluate the quality of our MT reference-free metrics by correlating them with human judgments of translation quality. These quality judgments are based on comparing human references and system predictions. We will discuss this discrepancy in §5.3.

Word-level metrics. We denote our word-level alignment metrics based on WMD as MoverScore-ngram + Align(Embedding), where Align is one of our two post-hoc cross-lingual alignment methods (CLP or UMD). For example, Mover-2 + UMD(M-BERT) denotes the metric combining MoverScore based on bigram alignments, with M-BERT embeddings and UMD as the post-hoc alignment method.

Sentence-level metric. We denote our sentence-level metrics as: Cosine + Align(Embedding). For example, Cosine + CLP(LASER) measures the cosine distance between the sentence embeddings obtained with LASER, post-hoc aligned with CLP.

4.1 Datasets

We collect the source language sentences, their system and reference translations from the WMT17-19 news translation shared task (Bojar et al., 2017b; Ma et al., 2018b, 2019), which contains predictions of 166 translation systems across 16 language pairs in WMT17, 149 translation systems across 14 language pairs in WMT18 and 233 translation systems across 18 language pairs in WMT19. We evaluate for X-en language pairs, selecting X from a set of 12 diverse languages: German (de), Chinese (zh), Czech (cs), Latvian (lv), Finnish (fi), Russian (ru), and Turkish (tr), Gujarati (gu), Kazakh (kk), Lithuanian (lt) and Estonian (et). Each language pair in WMT17-19 has approximately 3,000 source sentences, each associated to one reference translation and to the automatic translations generated by participating systems.

4.2 Baselines

We compare with a range of reference-free metrics: ibm1-morpheme and ibm1-pos4gram (Popović, 2012), LASIM (Yankovskaya et al., 2019), LP (Yankovskaya et al., 2019), YiSi-2 and YiSi-2-srl (Lo, 2019), and reference-based baselines BLEU (Papineni et al., 2002), SentBLEU (Koehn et al., 2007) and ChrF++ (Popović, 2017) for MT evaluation (see §2).[6] The main results are reported on WMT17. We report the results obtained on WMT18 and WMT19 in the Appendix.

[5] We linearly combine the cross-lingual metrics with the LM scores using a coefficient of 0.1 for all setups. We choose this value based on initial experiments on one language pair.

[6] The code of these unsupervised metrics is not released, thus we compare to their official results on WMT19 only.

Setting	Metrics	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
$m(\mathbf{y}^*, \mathbf{y})$	SENTBLEU	43.5	43.2	57.1	39.3	48.4	53.8	51.2	48.1
	CHRF++	**52.3**	**53.4**	**67.8**	**52.0**	**58.8**	**61.4**	**59.3**	**57.9**
$m(\mathbf{x}, \mathbf{y})$	*Baseline with Original Embeddings*								
	MOVER-1 + M-BERT	22.7	37.1	34.8	26.0	26.7	**42.5**	**48.2**	34.0
	COSINE + LASER	**32.6**	**40.2**	**41.4**	**48.3**	**36.3**	42.3	46.7	**41.1**
	Cross-lingual Alignment for Sentence Embedding								
	COSINE + CLP(LASER)	33.4	**40.5**	42.0	**48.6**	**36.0**	44.7	42.2	**41.1**
	COSINE + UMD(LASER)	**36.6**	28.1	**45.5**	48.5	31.3	**46.2**	**49.4**	40.8
	Cross-lingual Alignment for Word Embedding								
	MOVER-1 + RCSLS	18.9	26.4	31.9	33.1	25.7	31.1	34.3	28.8
	MOVER-1 + CLP(M-BERT)	33.4	38.6	50.8	48.0	33.9	**51.6**	53.2	44.2
	MOVER-2 + CLP(M-BERT)	**33.7**	38.8	**52.2**	**50.3**	**35.4**	51.0	**53.3**	**45.0**
	MOVER-1 + UMD(M-BERT)	22.3	38.1	34.5	30.5	31.2	43.5	48.6	35.5
	MOVER-2 + UMD(M-BERT)	23.1	38.9	37.1	34.7	33.0	44.8	48.9	37.2
	Combining Language Model								
	COSINE + CLP(LASER) ⊕ LM	48.8	46.7	63.2	66.2	51.0	54.6	48.6	54.2
	COSINE + UMD(LASER) ⊕ LM	**49.4**	46.2	**64.7**	66.4	**51.1**	**56.0**	52.8	**55.2**
	MOVER-2 + CLP(M-BERT) ⊕ LM	46.5	46.4	63.3	63.8	47.6	55.5	**53.5**	53.8
	MOVER-2 + UMD(M-BERT) ⊕ LM	41.8	**46.8**	60.4	59.8	46.1	53.8	52.4	51.6

Table 1: Pearson correlations with segment-level human judgments on the WMT17 dataset.

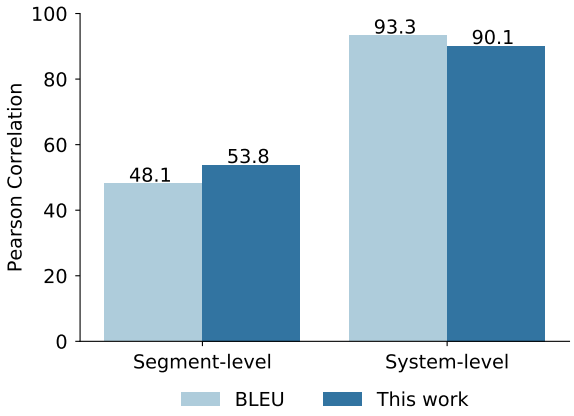

Figure 1: Average results of our best-performing metric, together with reference-based BLEU on WMT17.

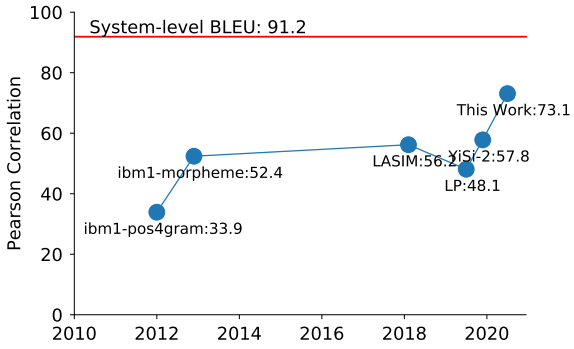

Figure 2: Average results of our metric best-performing metric, together with the official results of reference-free metrics, and reference-based BLEU on system-level WMT19.

4.3 Results

Figure 1 shows that our metric MOVER-2 + CLP(M-BERT) ⊕ LM, operating on modified M-BERT with the post-hoc re-mapping and combining a target-side LM, outperforms BLEU by 5.7 points in segment-level evaluation and achieves comparable performance in the system-level evaluation. Figure 2 shows that the same metric obtains 15.3 points gains (73.1 vs. 57.8), averaged over 7 languages, on WMT19 (system-level) compared to the the state-of-the-art reference-free metric YiSi-2. Except for one language pair, gu-en, our metric performs on a par with the reference-based BLEU (see Table 8 in the Appendix) on system-level.

In Table 1, we exhaustively compare results for several of our metric variants, based either on M-BERT or LASER. We note that re-mapping has considerable effect for M-BERT (up to 10 points improvements), but much less so for LASER. We believe that this is because the underlying embedding space of LASER is less 'misaligned' since it has been (pre-)trained on parallel data.[7] While the re-mapping is thus effective for metrics based on M-BERT, we still require the target-side LM to outperform BLEU. We assume the LM can address

[7]However, in the appendix, we find that re-mapping LASER using 2k parallel sentences achieves considerable improvements on low-resource languages, e.g., kk-en (from -61.1 to 49.8) and lt-en (from 68.3 to 75.9); see Table 8.

challenges that the re-mapping apparently is not able to handle properly; see our discussion in §5.1.

Overall, we remark that none of our metric combinations performs consistently best. The reason may be that LASER and M-BERT are pretrained over hundreds of languages with substantial differences in corpora sizes in addition to the different effects of the re-mapping. However, we observe that MOVER-2 + CLP(M-BERT) performs best on average over all language pairs when the LM is not added. When the LM is added, MOVER-2 + CLP(M-BERT) $\oplus$ LM and COSINE + UMD (LASER) $\oplus$ LM perform comparably. This indicates that there may be a saturation effect when it comes to the LM or that the LM coefficients should be tuned individually for each semantic similarity metric based on cross-lingual representations.

5 Analysis

We first analyze preferences of our metrics based on M-BERT and LASER (§5.1) and then examine how much parallel data we need for re-mapping our vector spaces (§5.2). Finally, we discuss whether it is legitimate to correlate our metric scores, which evaluate the similarity of system predictions and source texts, to human judgments based on system predictions and references (§5.3).

5.1 Metric preferences

To analyze why our metrics based on M-BERT and LASER perform so badly for the task of reference-free MT evaluation, we query them for their preferences. In particular, for a fixed source sentence $\mathbf{x}$, we consider two target sentences $\tilde{\mathbf{y}}$ and $\hat{\mathbf{y}}$ and evaluate the following score difference:

$$d(\tilde{\mathbf{y}}, \hat{\mathbf{y}}; \mathbf{x}) := m(\mathbf{x}, \tilde{\mathbf{y}}) - m(\mathbf{x}, \hat{\mathbf{y}}) \qquad (1)$$

When $d > 0$, then metric m prefers $\tilde{\mathbf{y}}$ over $\hat{\mathbf{y}}$, given $\mathbf{x}$, and when $d < 0$, this relationship is reversed. In the following, we compare preferences of our metrics for specifically modified target sentences $\tilde{\mathbf{y}}$ over the human references $\mathbf{y}^\star$. We choose $\tilde{\mathbf{y}}$ to be (i) a random reordering of $\mathbf{y}^\star$, to ensure that our metrics do not have the BOW (bag-of-words) property, (ii) a word-order preserving translation of $\mathbf{x}$, i.e., (ii-a) an expert reordering of the human $\mathbf{y}^\star$ to have the same word order as $\mathbf{x}$ as well as (ii-b) a word-by-word translation, obtained either using experts or automatically. Especially condition (ii-b) tests for preferences for literal translations, a common MT-system property.

Expert word-by-word translations. We had an expert (one of the co-authors) translate 50 German sentences word-by-word into English. Table 2 illustrates this scenario. We note how bad the word-by-word translations sometimes are even for closely related language pairs such as German-English. For example, the word-by-word translations in English retain the original German verb final positions, leading to quite ungrammatical English translations.

Figure 3 shows histograms for the d statistic for the 50 selected sentences. We first check condition (i) for the 50 sentences. We observe that both MOVER + M-BERT and COSINE+LASER prefer the original human references over random reorderings, indicating that they are not BOW models, a reassuring finding. Concerning (ii-a), they are largely indifferent between correct English word order and the situation where the word order of the human reference is the same as the German. Finally, they strongly prefer the expert word-by-word translations over the human references (ii-b).

Condition (ii-a) in part explains why our metrics prefer expert word-by-word translations the most: for a given source text, these have higher lexical overlap than human references and, by (ii-a), they have a favorable target language syntax, *viz.*, where the source and target language word order are equal. Preference for translationese, (ii-b), in turn is apparently a main reason why our metrics do not perform well, by themselves and without a language model, as reference-free MT evaluation metrics. More worryingly, it indicates that cross-lingual M-BERT and LASER are not robust to the 'adversarial inputs' given by MT systems.

Automatic word-by-word translations. For a large-scale analysis of condition (ii-b) across different language pairs, we resort to automatic word-by-word translations obtained from Google Translate (GT). To do so, we go over each word in the source sentence $\mathbf{x}$ from left to right, look up its translation in GT *independently of context* and replace the word by the obtained translation. When a word has several translations, we keep the first one offered by GT. Due to context-independence, the GT word-by-word translations are of much lower quality than the expert word-by-word translations since they often pick the wrong word senses—e.g., the German word *sein* may either be a personal pronoun (*his*) or the infinitive *to be*, which would be selected correctly only by chance; cf. Table 2.

$\mathbf{x}$	Dieser von Langsamkeit geprägte Lebensstil scheint aber ein Patentrezept für ein hohes Alter zu sein.
$\mathbf{y}^\star$	However, this slow pace of life seems to be the key to a long life.
$\mathbf{y}^\star$-random	To pace slow seems be the this life. life to a key however, of long
$\mathbf{y}^\star$-reordered	This slow pace of life seems however the key to a long life to be.
$\mathbf{x}'$-GT	This from slowness embossed lifestyle seems but on nostrum for on high older to his.
$\mathbf{x}'$-expert	This of slow pace characterized life style seems however a patent recipe for a high age to be.
$\mathbf{x}$	Putin teilte aus und beschuldigte Ankara, Russland in den Rücken gefallen zu sein.
$\mathbf{y}^\star$	Mr Putin lashed out, accusing Ankara of stabbing Moscow in the back.
$\mathbf{y}^\star$-random	Moscow accusing lashed Putin the in Ankara out, Mr of back. stabbing
$\mathbf{y}^\star$-reordered	Mr Putin lashed out, accusing Ankara of Moscow in the back stabbing.
$\mathbf{x}'$-GT	Putin divided out and accused Ankara Russia in the move like to his.
$\mathbf{x}'$-expert	Putin lashed out and accused Ankara, Russia in the back fallen to be.

Table 2: Original German input sentence $\mathbf{x}$, together with the human reference $\mathbf{y}^\star$, in English, and a randomly ($\mathbf{y}^\star$-random) and expertly reordered ($\mathbf{y}^\star$-reordered) English sentence as well as expert word-by-word translation ($\mathbf{x}'$) of the German source sentence. The latter is either obtained by the human expert or by Google Translate (GT).

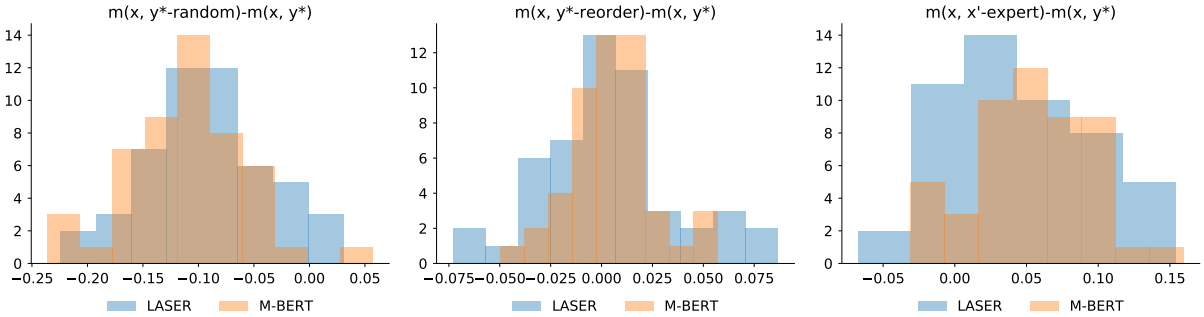

Figure 3: Histograms of d scores defined in Eq. (1). Left: Metrics based on LASER and M-BERT favor gold over randomly-shuffled human references. Middle: Metrics are roughly indifferent between gold and reordered human references. Right: Metrics favor expert word-by-word translations over gold human references.

Instead of reporting histograms of d, we define a "W2W" statistic that counts the relative number of times that $d(\mathbf{x}', \mathbf{y}^\star)$ is positive, where $\mathbf{x}'$ denotes the described literal translation of $\mathbf{x}$ into the target language:

$$\text{W2W} := \frac{1}{N} \sum_{(\mathbf{x}', \mathbf{y}^\star)} I(\, d(\mathbf{x}', \mathbf{y}^\star) > 0\,) \qquad (2)$$

Here N normalizes W2W to lie in $[0, 1]$ and a high W2W score indicates the metric prefers translationese over human-written references. Table 3 shows that reference-free metrics with original embeddings (LASER and M-BERT) either still prefer literal over human translations (e.g., W2W score of 70.2% for cs-en) or struggle in distinguishing them. Re-mapping helps to a small degree. Only when combined with the LM scores do we get adequate scores for the W2W statistic. Indeed, the LM is expected to capture unnatural word order in the target language and penalize word-by-word translations by recognizing them as much less likely to appear in a language.

Note that for expert word-by-word translations, we would expect the metrics to perform even worse.

Metrics	cs-en	de-en	fi-en
COSINE + LASER	70.2	65.7	53.9
COSINE + CLP(LASER)	70.7	64.8	53.7
COSINE + UMD(LASER)	67.5	59.5	52.9
COSINE + UMD(LASER) $\oplus$ LM	**7.0**	**7.1**	**6.4**
MOVER-2 + M-BERT	61.8	50.2	45.9
MOVER-2 + CLP(M-BERT)	44.6	44.5	32.0
MOVER-2 + UMD(M-BERT)	54.5	44.3	39.6
MOVER-2 + CLP(M-BERT) $\oplus$ LM	**7.3**	**10.2**	**6.4**

Table 3: W2W statistics for selected language pairs. Numbers are in percent.

5.2 Size of Parallel Corpora

Figure 4 compares sentence- and word-level re-mapping trained with a varying number of parallel sentences. Metrics based on M-BERT result in the highest correlations after re-mapping, even with a small amount of training data (1k). We observe that COSINE + CLP(LASER) and MOVER-2 + CLP(M-BERT) show very similar trends with a sharp increase with increasing amounts of parallel data and then level off quickly. However, the M-BERT based Mover-2 reaches its peak and outperforms the original baseline with only 1k data, while LASER needs 2k before beating the corre-

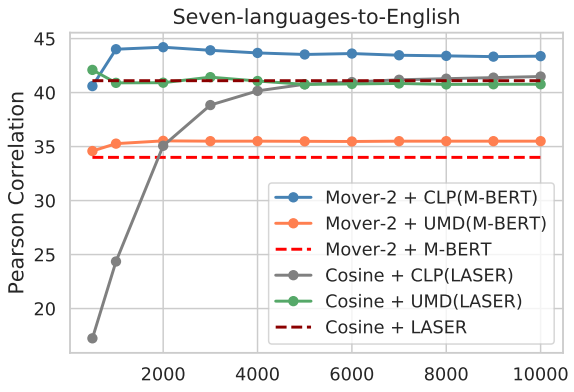

Figure 4: Average results of our metrics based on sentence- and word-based re-mappings of vector spaces as a function of different sizes of parallel corpus (x-axis).

sponding original baseline.

5.3 Human Judgments

The WMT datasets contain segment- and system-level human judgments that we use for evaluating the quality of our reference-free metrics. The segment-level judgments assign one direct assessment (DA) score to each pair of system and human translation, while system-level judgments associate each system with a single DA score averaged across all pairs in the dataset. We initially suspected the DA scores to be biased for our setup—which compares $\mathbf{x}$ with $\mathbf{y}$—as they are based on comparing $\mathbf{y}^{\star}$ and $\mathbf{y}$. Indeed, it is known that (especially) human professional translators "improve" $\mathbf{y}^{\star}$, e.g., by making it more readable, relative to the original $\mathbf{x}$ (Rabinovich et al., 2017). We investigated the validity of DA scores by collecting human assessments in the cross-lingual settings (CLDA), where annotators directly compare source and translation pairs $(\mathbf{x}, \mathbf{y})$ from the WMT17 dataset. This small-scale manual analysis hints that DA scores are a valid proxy for CLDA. Therefore, we decided to treat them as reliable scores for our setup and evaluate our proposed metrics by comparing their correlation with DA scores.

6 Conclusion

Existing semantically-motivated metrics for reference-free evaluation of MT systems have so far displayed rather poor correlation with human estimates of translation quality. In this work, we investigate a range of reference-free metrics based on cutting-edge models for inducing cross-lingual semantic representations: cross-lingual (contextualized) word embeddings and cross-lingual sentence embeddings. We have identified some scenarios in which these metrics fail, prominently their inability to punish literal word-by-word translations (the so-called "translationese"). We have investigated two different mechanisms for mitigating this undesired phenomenon: (1) an additional (weakly-supervised) cross-lingual alignment step, reducing the mismatch between representations of mutual translations, and (2) language modeling (LM) on the target side, which is inherently equipped to punish "unnatural" sentences in the target language. We show that the reference-free coupling of cross-lingual similarity scores with the target-side language model surpasses the reference-based BLEU in segment-level MT evaluation.

We believe our results have two relevant implications. First, they portray the viability of reference-free MT evaluation and warrant wider research efforts in this direction. Second, they indicate that reference-free MT evaluation may be the most challenging ("adversarial") evaluation task for multilingual text encoders as it uncovers some of their shortcomings—prominently, the inability to capture semantically non-sensical word-by-word translations or paraphrases—which remain hidden in their common evaluation scenarios.

We release our metrics under the name *XMoverScore* publicly: https://github.com/AIPHES/ACL20-Reference-Free-MT-Evaluation.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions, which greatly improved the final version of the paper. This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1. The contribution of Goran Glavaš is supported by the Eliteprogramm of the Baden-Württemberg-Stiftung, within the scope of the grant AGREE.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego,

California. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Ondrej Bojar, Yvette Graham, and Amir Kamran. 2017a. Results of the WMT17 metrics shared task. In *Proceedings of the Conference on Machine Translation (WMT)*.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017b. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 879–887.

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. 2002. 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, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2019. On the evaluation of machine translation systems trained with back-translation. *CoRR*, abs/1908.05204.

Marina Fomicheva and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016-Short Papers*, pages 77–82. ACL Home Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of ACL*.

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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vigas, 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.

Martin Josifoski, Ivan S Paskov, Hristo S Paskov, Martin Jaggi, and Robert West. 2019. Crosslingual document embedding as reduced-rank ridge regression. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 744–752.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2019. A general framework for implicit and explicit debiasing of distributional word vector spaces. *arXiv preprint arXiv:1909.06092*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

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, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. XMEANT: Better semantic MT evaluation without reference translations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 765–771, Baltimore, Maryland. Association for Computational Linguistics.

Qingsong Ma, Ondrej Bojar, and Yvette Graham. 2018a. Results of the WMT18 metrics shared task. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018b. 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, 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.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Bari Saiful M Mohiuddin, Tasnim and Shafiq Joty. 2020. Lnmap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. *CoRR, abs/1309.4168*.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2012. Morpheme- and POS-based IBM1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 133–137, Montréal, Canada. 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.

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.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. 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 (WMT)*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1181. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? 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 4398–4409.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings*

of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.

7 Appendix

7.1 Zero-shot Transfer to Resource-lean Language

Our metric allows for estimating translation quality on new domains. However, the evaluation is limited to those languages covered by multilingual embeddings. This is a major drawback for low-resource languages—e.g., Gujarati is not included in LASER. To this end, we take multilingual USE (Yang et al., 2019) as an illustrating example which covers only 16 languages (in our sample Czech, Latvian and Finish are not included in USE). We re-align the corresponding embedding spaces with our re-mapping functions to induce evaluation metrics even for these languages, using only 2k translation pairs. Table 4 shows that our metric with a composition of re-mapping functions can raise correlation from zero to 0.10 for cs-en and to 0.18 for lv-en. However, for one language pair, fi-en, we see correlation goes from negative to zero, indicating that this approach does not always work. This observation warrants further investigation.

Metrics	cs-en	fi-en	lv-en
BLEU	0.849	0.834	0.946
COSINE + LAS	-0.001	-0.149	0.019
COSINE + CLP(USE)	0.072	-0.068	0.109
COSINE + UMD(USE)	0.056	-0.061	0.113
COSINE + CLP ∘ UMD(USE)	0.089	-0.030	0.162
COSINE + UMD ∘ CLP(USE)	**0.102**	-0.007	**0.180**

Table 4: The Pearson correlation of merics on segment-level WMT17. '∘' marks the composition of two re-mapping functions.

Setting	Metrics	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
$m(\mathbf{y}^*, \mathbf{y})$	BLEU	**0.971**	0.923	0.903	**0.979**	0.912	**0.976**	0.864	0.933
	CHRF++	0.940	**0.965**	**0.927**	0.973	**0.945**	0.960	**0.880**	**0.941**
$m(\mathbf{x}, \mathbf{y})$	*Baseline with Original Embeddings*								
	MOVER-1 + M-BERT	0.408	**0.905**	0.570	0.571	0.855	0.576	**0.816**	0.672
	COSINE + LASER	**0.821**	0.821	**0.744**	**0.754**	**0.895**	**0.890**	0.676	**0.800**
	Cross-lingual Alignment for Sentence Embedding								
	COSINE + CLP(LASER)	0.824	0.830	**0.760**	**0.766**	0.900	**0.942**	**0.757**	**0.826**
	COSINE + UMD(LASER)	**0.833**	**0.858**	0.735	0.754	**0.909**	0.870	0.630	0.798
	Cross-lingual Alignment for Word Embedding								
	MOVER-1 + RCSLS	-0.693	-0.053	0.738	0.251	0.538	0.380	0.439	0.229
	MOVER-1 + CLP(M-BERT)	0.796	0.960	0.879	0.874	0.894	0.864	0.898	0.881
	MOVER-2 + CLP(M-BERT)	**0.818**	0.971	**0.885**	**0.887**	0.878	**0.893**	0.896	**0.890**
	MOVER-1 + UMD(M-BERT)	0.610	0.956	0.526	0.599	**0.906**	0.538	0.898	0.719
	MOVER-2 + UMD(M-BERT)	0.650	**0.973**	0.574	0.649	0.888	0.634	**0.901**	0.753
	Combining Language Model								
	COSINE + CLP(LASER) $\oplus$ LM	**0.986**	0.909	0.868	0.968	0.858	0.910	0.800	0.900
	COSINE + UMD(LASER) $\oplus$ LM	0.984	0.904	0.861	**0.968**	0.850	**0.922**	**0.817**	**0.901**
	MOVER-2 + CLP(M-BERT) $\oplus$ LM	0.977	0.923	**0.873**	0.944	0.863	0.880	0.803	0.895
	MOVER-2 + UMD(M-BERT) $\oplus$ LM	0.968	**0.934**	0.832	0.951	**0.871**	0.862	0.821	0.891

Table 5: Pearson correlations with system-level human judgments on the WMT17 dataset.

Setting	Metrics	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	Average
$m(\mathbf{y}^*, \mathbf{y})$	SENTBLEU	0.233	0.415	0.285	0.154	0.228	0.145	0.178	0.234
	YISI-1	**0.319**	**0.488**	**0.351**	**0.231**	**0.300**	**0.234**	**0.211**	**0.305**
$m(\mathbf{x}, \mathbf{y})$	*Baseline with Original Embeddings*								
	MOVER-1 + M-BERT	0.005	0.229	0.179	0.115	0.100	0.039	**0.082**	0.107
	COSINE + LASER	**0.072**	**0.317**	**0.254**	**0.155**	**0.102**	**0.086**	0.064	**0.150**
	Cross-lingual Alignment for Word Embedding								
	COSINE + CLP(LASER)	0.093	0.323	0.254	**0.151**	0.112	0.086	0.074	0.156
	COSINE + UMD(LASER)	0.077	0.317	0.252	**0.145**	0.136	0.083	0.053	0.152
	COSINE + UMD $\circ$ CLP(LASER)	0.090	**0.337**	**0.255**	0.139	**0.145**	**0.090**	**0.088**	**0.163**
	COSINE + CLP $\circ$ UMD(LASER)	**0.096**	0.331	0.254	0.153	0.122	0.084	0.076	0.159
	Cross-lingual Alignment for Sentence Embedding								
	MOVER-1 + CLP(M-BERT)	**0.084**	0.279	0.207	0.147	**0.145**	**0.089**	**0.122**	**0.153**
	MOVER-2 + CLP(M-BERT)	0.063	**0.283**	0.193	0.149	0.136	0.069	0.115	0.144
	MOVER-1 + UMD(M-BERT)	0.043	0.264	0.193	0.136	0.138	0.051	0.113	0.134
	MOVER-2 + UMD(M-BERT)	0.040	0.268	0.188	0.143	0.141	0.055	0.111	0.135
	MOVER-1 + UMD $\circ$ CLP(M-BERT)	0.024	0.282	0.192	0.144	0.133	0.085	0.089	0.136
	MOVER-1 + CLP $\circ$ UMD(M-BERT)	0.073	0.277	**0.208**	0.148	0.142	0.086	0.121	0.151
	MOVER-2 + CLP $\circ$ UMD(M-BERT)	0.057	0.283	0.194	**0.149**	0.137	0.069	0.114	0.143
	Combining Language Model								
	COSINE + UMD $\circ$ CLP(LASER) $\oplus$ LM	0.288	0.455	0.226	0.321	0.263	0.159	0.192	0.272
	COSINE + CLP $\circ$ UMD(LASER) $\oplus$ LM	0.283	0.457	0.228	0.321	0.265	0.150	0.198	0.272
	MOVER-1 + CLP $\circ$ UMD(M-BERT) $\oplus$ LM	0.268	0.428	0.292	0.213	0.261	0.152	0.192	0.258
	MOVER-2 + CLP $\circ$ UMD(M-BERT) $\oplus$ LM	0.254	0.426	0.285	0.203	0.251	0.146	0.193	0.251

Table 6: Pearson correlations with segment-level human judgments on the WMT18 dataset.

Setting	Metrics	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	Average
$m(\mathbf{y}^*, \mathbf{y})$	BLEU	**0.970**	0.971	**0.986**	**0.973**	0.979	0.657	**0.978**	0.931
	METEOR++	0.945	**0.991**	0.978	0.971	**0.995**	**0.864**	0.962	**0.958**
$m(\mathbf{x}, \mathbf{y})$	*Baseline with Original Embeddings*								
	MOVER-1 + M-BERT	-0.629	0.915	0.880	0.804	0.847	0.731	**0.677**	0.604
	COSINE + LASER	**-0.348**	**0.932**	**0.930**	**0.906**	**0.902**	**0.832**	0.471	**0.661**
	Cross-lingual Alignment for Sentence Embedding								
	COSINE + CLP(LASER)	-0.305	0.934	0.937	0.908	0.904	0.801	0.634	0.688
	COSINE + UMD(LASER)	-0.241	0.944	0.933	0.906	0.902	0.842	0.359	0.664
	COSINE + UMD ∘ CLP(LASER)	**0.195**	**0.955**	**0.958**	**0.913**	0.896	**0.899**	**0.784**	**0.800**
	COSINE + CLP ∘ UMD(LASER)	-0.252	0.942	0.941	0.908	**0.919**	0.811	0.642	0.702
	Cross-lingual Alignment for Word Embedding								
	MOVER-1 + CLP(M-BERT)	**-0.163**	0.943	**0.918**	**0.941**	0.915	0.628	**0.875**	**0.722**
	MOVER-2 + CLP(M-BERT)	-0.517	0.944	0.909	0.938	0.913	0.526	0.868	0.654
	MOVER-1 + UMD(M-BERT)	-0.380	0.927	0.897	0.886	0.919	**0.679**	0.855	0.683
	MOVER-2 + UMD(M-BERT)	-0.679	0.929	0.891	0.896	**0.920**	0.616	0.858	0.633
	MOVER-1 + UMD ∘ CLP(M-BERT)	-0.348	**0.949**	0.905	0.890	0.905	0.636	0.776	0.673
	MOVER-1 + CLP ∘ UMD(M-BERT)	-0.205	0.943	0.916	0.938	0.913	0.641	0.871	0.717
	MOVER-2 + CLP ∘ UMD(M-BERT)	-0.555	0.944	0.908	0.935	0.911	0.551	0.863	0.651
	Combining Language Model								
	COSINE + UMD ∘ CLP(LASER) ⊕ LM	0.979	0.967	0.979	0.947	0.942	0.673	0.954	0.919
	COSINE + CLP ∘ UMD(LASER) ⊕ LM	0.974	0.966	0.983	0.951	0.951	0.255	0.961	0.863
	MOVER-1 + CLP ∘ UMD(M-BERT) ⊕ LM	0.956	0.960	0.949	0.973	0.951	0.097	0.954	0.834
	MOVER-2 + CLP ∘ UMD(M-BERT) ⊕ LM	0.959	0.961	0.947	0.979	0.951	-0.036	0.952	0.815

Table 7: Pearson correlations with system-level human judgments on the WMT18 dataset.

		Direct Assessment							
Setting	Metrics	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	Average
$m(\mathbf{y}^*, \mathbf{y})$	BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899	0.907
$m(\mathbf{x}, \mathbf{y})$	*Existing Reference-free Metrics*								
	IBM1-MORPHEME(Popović, 2012)	0.345	0.740	-	-	0.487	-	-	-
	IBM1-POS4GRAM(Popović, 2012)	0.339	-	-	-	-	-	-	-
	LASIM(Yankovskaya et al., 2019)	0.247	-	-	-	-	0.310	-	-
	LP(Yankovskaya et al., 2019)	0.474	-	-	-	-	0.488	-	-
	YISI-2(Lo, 2019)	0.796	0.642	0.566	0.324	0.442	0.339	0.940	0.578
	YISI-2-SRL(Lo, 2019)	0.804	-	-	-	-	-	0.947	-
	Baseline with Original Embeddings								
	MOVER-1 + M-BERT	**0.358**	0.611	**-0.396**	**0.335**	**0.559**	**0.261**	**0.880**	**0.373**
	COSINE + LASER	0.217	**0.891**	-0.745	-0.611	0.683	-0.303	0.842	0.139
	Our Cross-lingual based Metrics								
	MOVER-2 + CLP(M-BERT)	**0.625**	0.890	-0.060	**0.993**	**0.851**	**0.928**	**0.968**	**0.742**
	COSINE + CLP(LASER)	0.225	**0.894**	**0.041**	0.150	0.696	-0.184	0.845	0.381
	COSINE + UMD ∘ CLP(LASER)	0.074	0.835	-0.633	**0.498**	**0.759**	-0.201	0.610	0.277
	Our Cross-lingual based Metrics ⊕ LM								
	COSINE + CLP(LASER) ⊕ LM	0.813	0.910	-0.070	-0.735	0.931	0.630	0.711	0.456
	COSINE + UMD(LASER) ⊕ LM	0.817	0.908	-0.383	-0.902	0.929	0.573	0.781	0.389
	MOVER-2 + CLP(M-BERT) ⊕ LM	0.848	0.907	**-0.068**	**0.775**	0.963	**0.866**	0.827	**0.731**
	MOVER-2 + UMD(M-BERT) ⊕ LM	**0.859**	**0.914**	-0.181	**-0.391**	**0.970**	0.702	**0.874**	0.535

Table 8: Pearson correlations with system-level human judgments on the WMT19 dataset. '-' marks the numbers not officially reported in (Ma et al., 2019).