Global Explainability of BERT-Based Evaluation Metrics by Disentangling along Linguistic Factors

Marvin Kaster, Wei Zhao, Steffen Eger
Natural Language Learning Group (NLLG)
Technische Universität Darmstadt, Germany
`marvin.kaster@stud.tu-darmstadt.de`
`{zhao,eger}@aiphes.tu-darmstadt.de`

Abstract

Evaluation metrics are a key ingredient for progress of text generation systems. In recent years, several BERT-based evaluation metrics have been proposed (including BERTScore, MoverScore, BLEURT, etc.) which correlate much better with human assessment of text generation quality than BLEU or ROUGE, invented two decades ago. However, little is known what these metrics, which are based on black-box language model representations, actually capture (it is typically assumed they model semantic similarity). In this work, we use a simple regression based global explainability technique to disentangle metric scores along linguistic factors, including semantics, syntax, morphology, and lexical overlap. We show that the different metrics capture all aspects to some degree, but that they are all substantially sensitive to lexical overlap, just like BLEU and ROUGE. This exposes limitations of these novelly proposed metrics, which we also highlight in an adversarial test scenario.

1 Introduction

Evaluation metrics are a key ingredient in assessing the quality of text generation systems, be it machine translation, summarization, or conversational AI models. Traditional evaluation metrics in machine translation and summarization, BLEU and ROUGE (Papineni et al., 2002a; Lin, 2004), have measured lexical n-gram overlap between system prediction and a human reference. While simple and easy to understand, early on, limitations of such lexical overlap metrics have been recognized (Callison-Burch et al., 2006), e.g., in that they can only measure surface level similarity, and they are especially inadequate when it comes to assessing current high-quality text generation systems (Rei et al., 2020; Mathur et al., 2020; Marie et al., 2021).

Recently, a class of novel evaluation metrics based on BERT and its variants has been explored that correlates much better with human assessments

of translation quality. For example, BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), BLEURT (Sellam et al., 2020), XMoverScore (Zhao et al., 2020), and COMET (Rei et al., 2020) all use large-scale pretrained language models, but differ in whether they compare hypotheses to references, to source texts, to both, on the one hand, and whether they use human scores for supervision or not, on the other. Since these models all leverage large-scale language models which have pushed the state-of-the-art in many areas of NLP, their success comes with little surprise.

To better understand these novel metrics based on black-box language representations is a prerequisite for identifying their limitations, e.g., to adversarial inputs. For example, if an evaluation metric is sensitive to lexical overlap, it can be fooled by using the same words but in different order.

In this work, we fill the existing 'explainability gap' and introspect linguistic properties encoded in BERT-based evaluation metrics. Although there is already considerable work on introspecting and understanding BERT (see Rogers et al. (2020) for an overview), e.g., via probing (Tenney et al., 2019), analyzes by Hewitt and Liang (2019); Eger et al. (2020); Ravichander et al. (2021) indicate that probing results (based on supervision) are not always trustworthy. More importantly, the modern evaluation metrics sketched above rely on at least two factors: BERT (or its variants) and different aggregation schemes, such as Earth Mover Distance (Kusner et al., 2015; Zhao et al., 2019) or greedy alignment (Zhang et al., 2020), on top of BERT. Understanding BERT alone is thus not sufficient for explaining BERT-based evaluation metrics.

Here, we present a simple global explanation technique of BERT-based evaluation metrics which disentangles them on prominent linguistic factors, *viz.*, syntax, semantics, morphology and lexical overlap. We find that all metrics capture these linguistic aspects to certain (but differing) degrees,

and they are particularly sensitive to lexical overlap, which makes them prone to similar adversarial fooling (cf. Li et al., 2020; Eger et al., 2019; Keller et al., 2021) as BLEU-based lexical overlap metrics. Overall, our contributions are:

- We disentangle a multitude of current BERT-based evaluation metrics on four linguistic factors using linear regression.

- We show that all metrics are sensitive to all factors and especially to lexical overlap, as confirmed both by the linear regression and an adversarial experiment.

- Based on the insight that different metrics capture different linguistic factors to varying degrees, we ensemble metrics and identify an average improvement of between 8 and 13% for the most heterogeneous metrics.

2 Related work

Our work concerns reference-based and reference-free metrics, on the one hand, and model introspection (or 'explainability'), on the other.

Evaluation metrics for Natural Language Generation In the last few years, several strong performing evaluation metrics have been proposed, the majority of which is based on BERT and similar high-quality text representations. They can be differentiated along two dimensions: (i) the input arguments they take, and (ii) whether they are supervised or unsupervised. *Reference-based* metrics compare human references to system predictions. Popular metrics are BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), and BLEURT (Sellam et al., 2020). In contrast, *reference-free* metrics directly compare source texts to system predictions, thus they are more resource-lean. Popular examples are XMoverScore (Zhao et al., 2020), Yisi-2 (Lo, 2019), KoBE (Gekhman et al., 2020), and SentSim (Song et al., 2021). Rei et al. (2020) use all three information signals: source text, hypotheses and human references. There are also reference-free metrics outside the field of machine translation; for example, SUPERT (Gao et al., 2020) for summarization. *Supervised* metrics train on human sentence-level scores, e.g., Direct Assessment (DA) scores or post-editing effort (HTER) for MT. These metrics include BLEURT and COMET (Rei et al., 2020). In MT, most metrics from the so-called 'Quality Estimation' (QE) tasks are also supervised reference-free metrics, e.g., TransQuest (Ranasinghe et al., 2020) and BERGAMOT-LATTE (Fomicheva et al., 2020b). *Unsupervised* metrics require no such supervisory signal (e.g., MoverScore, BERTScore, XMoverScore, SentSim).

Model introspection There has been a recent surge in interest in explaining deep learning models. The techniques for explainability differ in whether they provide justification or information for model outputs on individual instances (*local explainability*) or focus on a model as a whole and disclose its internal structure (*global explainability*) (Danilevsky et al., 2020). Popular examples for local explainability are LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) that find features from the input (such as particular words) relevant to model outputs.

Concerning (global) interpretability of text representations, previous works (Adi et al., 2017; Conneau et al., 2018) introspect the properties encoded in vector representations through probing classifiers—trained on external data to perform a certain linguistic task, such as inducing the dependency tree depth from a text representation (of a sentence). Tenney et al. (2019) extend this idea by inspecting BERT representations layer-by-layer, and find that BERT captures more semantic information in its higher layers and more syntactic and morphological information in its lower layers. However, probing results are not always trustworthy due to the sensitivity to probing design choices, e.g., data size and classifier choices (Eger et al., 2020), and data artefacts (Ravichander et al., 2021). More importantly, evaluation metrics use BERT differently: some are supervised and others are unsupervised, some fine-tune BERT on semantic similarity datasets, and they generally differ on how they aggregate and compare BERT representations. This means to understand these metrics it does not suffice to understand BERT alone.

In our work, we disentangle BERT based evaluation metrics along linguistic factors as a form of global explainability of those metrics. This yields insights into which linguistic information signals specific metrics use, in general, and may expose their limitations.

3 Our approach

In our scenario, we consider different metrics m taking two arguments and assigning them a real-

valued score

$$\mathfrak{m} : (x, y) \mapsto s_{\mathfrak{m}} \in \mathbb{R}$$

where x and y are source text and hypothesis text, respectively (for so-called *reference-free* metrics) or alternatively x and y are reference and hypothesis text, respectively (for so-called *reference-based* metrics). The scores s that metrics assign to x and y can be considered the *similarity* between x and y or *adequacy* of y given x. In our experiments in Section 4, we will focus on machine translation (MT) as use case; it is arguably the most popular and prominent text generation task. Thus, x is either a (sentence-level) source text in one language and y the corresponding MT output, or x is the human reference for the original source text.

To better understand evaluation metrics, we decompose their scores $s_{\mathfrak{m}}$ along multiple linguistic factors. An example is outlined in Table 1.

We follow a long line of research in the applied sciences, and use a *linear model* to explain a *target variable* (in our case, the metric score), also called *response variable*, in terms of multiple *regressors*, also called *explanatory variables*. That is, we estimate the linear regression

$$\begin{aligned} \mathfrak{m}(x, y) = \alpha \cdot sem(x, y) + \beta \cdot syn(x, y) \\ + \gamma \cdot lex(x, y) + \delta \cdot morph(x, y) + \epsilon \end{aligned} \quad (1)$$

Here, $sem(x, y)$, $syn(x, y)$, $morph(x, y)$ and $lex(x, y)$ are scores which describe the semantic, syntactic, morphological, and lexical similarity between the two argument sentences. The real coefficients α, β, γ and δ are the regressors' weights, estimated from data. Finally, ϵ is an error term.

Linear regression assumes a linear relationship between the target variable and the regressors. It may fail when the relationship is non-linear, but its simple model structure provides interpretability: a larger positive coefficient means the respective regressor has higher positive impact on the target variable (fixing all other variables), a coefficient close to zero means no linear relationship, and a larger negative coefficient means an inverse linear relationship between regressor and target variable.

The coefficient of determination R^2 describes how well the regression model reflects the data. It is defined as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

where *SSE* denotes the sum of squared errors and *SST* denotes the sum of squared totals. They are defined as

$$\text{SSE} = \sum_i (y_i - \hat{y}_i)^2, \quad \text{SST} = \sum_i (y_i - \overline{y}_i)^2$$

respectively, where $\hat{y}_i$ is the prediction of the model, y_i is the true score, and $\overline{y}_i$ is the mean, $\overline{y}_i = \frac{1}{N} \sum_i^N y_i$. R^2 is 1 for a perfect fit, 0 if it always predicts the mean and negative if the model is worse than this baseline.

To ensure comparability of the different regressions, we normalize the scores of our regressors and the response variable with the z-normalization, i.e., subtracting the mean and dividing by the standard deviation, per variable. In the following, we define our regressors.

Semantic score (SEM) The semantic scores are provided by the datasets and were annotated by humans who rated e.g. the translation quality. See Section 4.2 for details.

Syntactic score (SYN) To measure the syntactic similarity of the argument sentences x and y, we compare their dependency trees. Both sentences are parsed by the Stanford dependency parser (Chen and Manning, 2014). Then, the tree edit distance (TED) (Bille, 2005) between the resulting trees is calculated. As an extension of string edit distance, TED measures how many operations are necessary to transform one tree into the other. Only the structures of the trees are considered in the calculation. We ignore the actual words.

We normalized the TED to ensure comparability between sentences with different lengths (Zhang and Shasha, 1989). The final score is calculated as

$$syn(x, y) = 1 - \frac{TED}{l_x + l_y}$$

where l_x and l_y are the lengths of the sentences. Figure 1 shows an example of the TED calculation for sentences in the same language (monolingual) and Figure 2 shows a cross-lingual example.

Lexical overlap score (LEX) We measure the lexical overlap between x and y by the BLEU score (Papineni et al., 2002b): BLEU_n calculates the precision based on how many n-grams of one sentence can be found in the other sentence. In the experiments below, we use BLEU_1. Using unigrams assures that word order is ignored. The simple precision count is modified so that identical words are only counted once.

Hypothesis (y)	Reference/Source (x)	SEM	SYN	LEX	MOR
It is a boy , likes to sport , but it cannot do it because of their very.	He is a boy, he likes sports but he can't take part because of his knee.	-1.57	0.98	-0.59	-0.87
Zwei Besatzungsmitglieder galten als vermisst.	Two crew members were regarded as missing.	0.83	0.99	0.46	-2.40

Table 1: Example setups with normalized semantic, syntactic, lexical overlap and morphological scores.
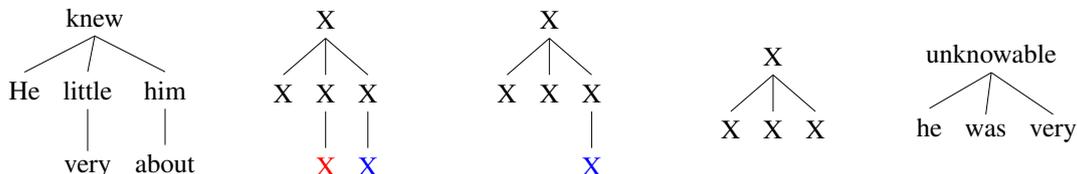

Figure 1: Monolingual tree edit distance example. The left-most tree is the first sentence of the sentence pair and the right-most tree is the second. To transform the left-most sentence into the right-most sentence, two leaves are removed. The unnormalized tree edit distance is thus 2. The normalized score is $1 - \frac{2}{6+4} = 0.8$.

For monolingual reference-based metrics, the BLEU score is calculated directly on the sentence pairs. To use BLEU for cross-lingual reference-free metrics, we choose to translate the non-English sentences into English via Google Translate, as it remains unclear how else to define lexical overlap between sentences from different languages. We compute BLEU scores on original English and translated sentence pairs.

Word 1	Word 2	UD Tags
reached	combined	Tense=Past VerbForm=Part
stay	sein	VerbForm=Inf

Table 2: Example of a morphological lexicon that contains word pairs with identical morphological UD tags, on which we finetune FastText word embeddings.

Morphological score (MOR) We introduce a morphological score $morph(x, y)$. To do so, we use static FastText word embeddings (Bojanowski et al., 2017) and increase morphological information in the original embeddings: (a) first, we produce two morphological lexicons based on words from WMT and STS, each containing word pairs with identical UD morphological tags (Nivre et al., 2020) (see Table 2). (b) Then, we finetune/retrofit the embeddings on the morphological lexicons using the method described in Faruqui et al. (2015), so that words with the same morphological tags have more similar representations.

The final morphological score for a sentence pair is the cosine similarity between two averaged sentence embeddings over refined word vectors

of each sentence. Note that, while we refer to these embeddings as morphological, they actually capture multiple linguistic factors and can only be considered *more* morphological than standard static vector spaces.

If the overlap of morphological features between a language pair is very low, the morphological score will not be meaningful. We exclude the morphological score for such language pairs.

4 Experiments

We analyze different evaluation metrics by calculating their score for sentence pairs. We use both reference-based metrics, which operate in a monolingual space, and reference-free metrics, which operate in a cross-lingual space.

4.1 Metrics

Reference-based metrics We consider the following reference-based metrics.

- BERTScore (Zhang et al., 2020) aggregates and compares BERT embeddings by determining a greedy alignment between words in two sentences and summing up the cosine similarities of representations of aligned words.

- MoverScore (Zhao et al., 2019) computes an optimal alignment between words in the two sentences using word mover distance (Kusner et al., 2015) on top of BERT representations.

- Sentence BERT (SBERT) (Reimers and Gurevych, 2019) fine-tunes Siamese BERT networks on NLI data, and produces sentence embeddings by using pooling on top of BERT
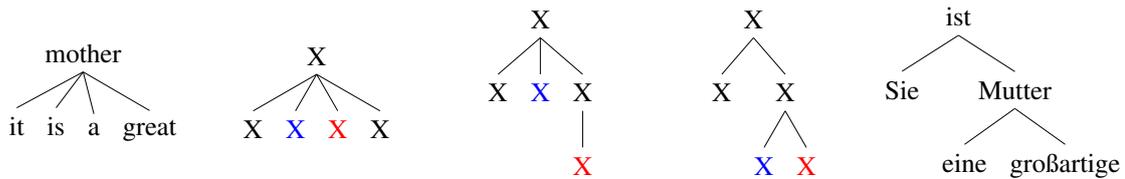
Figure 2: Cross-lingual tree edit distance example. To transform the left-most into the right-most sentence, two leaves (shown in red and blue) are moved to other locations. This takes 2 operations. The normalized score is $1 - \frac{2}{5+5} = 0.8$.

representations. We compute the cosine similarity between SBERT representations.

- SBERT-WK (Wang and Kuo, 2020) is a variant of SBERT which weighs different layers of BERT.

- In contrast to the others, BLEURT (Sellam et al., 2020) is a supervised metric and fine-tunes BERT on the WMT datasets with available human assessment of translation quality.

Reference-free metrics We consider the following reference-free metrics:

- Multilingual Sentence BERT (mSBERT) (Reimers and Gurevych, 2020) is a multilingual version of Sentence BERT trained on parallel data with multilingual knowledge distillation. A teacher model trained on STS and NLI provides English sentence embeddings. mSBERT (student model) is trained to produce embeddings for the English sentence and its translation which are close to the embeddings of the teacher model.

- Multilingual Universal Sentence Encoder (MUSE) (Yang et al., 2019) is a multilingual sentence embedding. It is a dual-encoder model which was trained on multiple tasks such as NLI and translation ranking. MUSE was trained on mono- and multilingual data.

- LaBSE (Feng et al., 2020) is a dual-encoder framework. The encoders re-use pre-trained BERT and finetune it with Masked Language Modeling and Translation Language modeling on monolingual and parallel data.

- LASER (Artetxe and Schwenk, 2019) is a BiLSTM encoder trained on parallel corpora. It produces language-agnostic representations. The encoder-decoder architecture is trained jointly on different languages.

- XMoverScore (Zhao et al., 2020) extends MoverScore to operate in the cross-lingual setup, and relies on re-aligned multilingual BERT representations. Note that we exclude a target-side lan-

guage model integrated in XMoverScore to have a similar setup as for MoverScore.

Except for XMoverScore, all metrics are based on calculation of cosine similarity between the source-language and target-language sentence embeddings. Except for MUSE and LASER, all metrics are based on BERT representations. Note that some multilingual reference-free metrics can also be used in the monolingual reference-based case, especially those based on calculating cosine similarity on top of sentence embeddings, thus we will include them in both settings.

4.2 Datasets

We use the datasets of the WMT shared task and the Semantic Text Similarity Benchmark (STSB) in our experiments. In the appendix, Table 8 shows the statistics in each dataset, and Table 9 shows examples for the sentences of the dataets.

WMT The WMT datasets contain an input sentence in the source language, the hypothesis translation of an MT system and a human reference sentence in the target language. Humans have rated the similarity between human reference and MT hypothesis using so-called 'direct assessment' (DA) scores which are framed in terms of one sentence 'adequately expressing the meaning' of another. We use these ratings as semantic scores in our setup.

For the reference-based case, we use the hypotheses and the references as sentence pairs. This data is collected over multiple language pairs which have English as target language (so both the human reference and the hypothesis are in English) of WMT15-WMT17. For the reference-free scenario, we pair the source texts with MT hypotheses and use the corresponding reference-to-hypothesis DA scores for similarity score. For German, we take the data from WMT15 (Bojar et al., 2015), WMT16 (Bojar et al., 2016) and WMT17 (Bojar et al., 2017). Chinese is only available in WMT17.

STSB The Semantic Text Similarity Benchmark (Cer et al., 2017) consists of English sentence pairs and a semantic similarity score for each pair. The scores were annotated by humans. The score is used as semantic score in the regression. In contrast to WMT, some sentence pairs in STSB are designed to have a different structure but a similar meaning.

While we use the sentence pairs directly for the monolingual case, we translate the sentences in the cross-lingual case using Google Translate, following Chidambaram et al. (2018).

5 Results

Reference-based Metrics Table 3 shows the results for the reference-based metrics. The R^2 values range from 0.43 to 0.76 on WMT, and from 0.58 to 0.91 on STS. This means we can reasonably well explain the metrics from our four explanatory variables and using a linear model. mSBERT can best be explained with an R^2 value of 0.91 on STS; however, since it has been trained on STS, this merely indicates overfitting. All metrics have positive coefficients for SEM, indicating that they all reflect semantic similarity and (semantic) 'adequacy' (as measured by DA), respectively: the weights range from 0.19 to 0.48 on WMT and from 0.12 to 0.76 on STS (ignoring mSBERT). The SYN coefficients are much lower and range from -0.05 to 0.16 for WMT and from -0.03 to 0.24 on STS. Mover-Score and BERTScore are most affected by syntactic similarity (0.11 to 0.24), while the sentence embedding based techniques have coefficients around zero. The coefficients for MOR are low on STS, except for SBERT-WK, and moderate for WMT.

All metrics have comparatively large coefficients for lexical overlap, especially on WMT: the coefficient values range from 0.24 to 0.64 on WMT and from 0.14 to 0.67 on STS. Especially LEX dominates for MoverScore and BERTScore, indicating that these two metrics are most sensitive to lexical adversaries, potentially making them vulnerable to inputs such as 'man bites dog' vs. 'dog bites man'.

Reference-free Metrics Table 4 shows the results for ZH-EN in the reference-free setup (omitting the score for MOR as indicated earlier). Many SYN coefficients are now zero or negative, meaning that a larger syntactic difference between the input arguments leads to a higher metric score, indicating that metrics are sensitive to syntactic language differences. LEX is still significant in all cases. SEM has higher coefficient values than LEX

in 6 out of 10 cases, and when it 'wins', it wins by a large margin. However, we note that the R^2 are low: they range from 0.3-0.39 on WMT and from 0.24-0.59 on STS. This means we can either not (well) explain the metrics given our current regressors or the relationship is not well explained by a linear model. The results for DE-EN are similar; we provide them in Table 10 (appendix).

To explore why R^2 scores are now lower, we note that reference-free metrics based on BERT might contain a form of *cross-lingual bias* (CLB) in that they do not properly score mutual translations, as Cao et al. (2020) and many others have shown that the multilingual subspaces induced by BERT are mis-aligned. We thus include a factor CLB as a regressor to measure how significant this bias is in different metrics. Note that the metrics use either different BERT variants or other representations such as LASER, which points to different sources of CLB. Therefore, we realize CLB differently across metrics. For each metric regression, we use the same metric but take a parallel sentence, i.e., source text and Google translation (as we assume that Google Translate has very high quality in general), as input arguments, and take the metric score as a proxy of the CLB factor. If a metric does not contain cross-lingual bias, it should assign almost full scores to parallel sentences; this constant would then be meaningless in the regression.

In Table 5, we show that including the CLB factor in the regression improves the R^2. We substantially improve the R^2 for XMoverScore (from 0.39 to 0.68), but observe little improvements for the remaining metrics (especially on STS). This is because XMoverScore is more problematic than the other metrics in terms of properly scoring mutual translations, given that the other metrics use BERT variants (or LASER) that have been finetuned (or trained) on parallel sentences. Apart from CLB, both SEM and LEX are the dominating factors in the regression. The DE-EN results are similar—see Table 11 (appendix).

Limitations The R^2 scores for the WMT dataset are almost always lower than the corresponding STS scores. One may not forget that STS sentences are in a sense artificial sentences of the form 'a girl is playing a guitar' while WMT contains more realistic sentences as well as their (possibly faulty, non-grammatical) translations. The WMT datasets are furthermore inhomogeneous in that we used different years from 2015 to 2017, which has dif-

	WMT						STS				
Metric	SEM	SYN	LEX	MOR	R^2	**Metric**	SEM	SYN	LEX	MOR	R^2
MoverScore	0.28	0.15	0.64	-0.06	0.76	MoverScore	0.30	0.24	0.45	0.04	0.61
BERTScore	0.27	0.16	0.61	-0.01*	0.74	BERTScore	0.12	0.11	0.67	0.06	0.69
LASER	0.19	0.04	0.33	0.32	0.53	LASER	0.55	0.05	0.28	0.09	0.63
SBERT	0.37	-0.02*	0.22	0.22	0.42	SBERT	0.73	-0.03	0.14	-0.06	0.60
SBERT-WK	0.30	-0.02	0.32	0.33	0.58	SBERT-WK	0.26	0.04	0.31	0.30	0.55
LaBSE	0.31	0.00*	0.36	0.22	0.53	LaBSE	0.60	0.07	0.27	0.06	0.67
mSBERT	0.41	-0.05	0.24	0.18	0.43	mSBERT	0.93	-0.02	0.04	0.00*	0.91
BLEURT	0.48	0.07	0.30	0.05	0.57	BLEURT	0.62	0.08	0.19	0.03*	0.58
mUSE	0.27	-0.03	0.39	0.19	0.49	mUSE	0.76	0.00*	0.22	-0.11	0.68

Table 3: Regression results for reference-based metrics. Coefficient values for linguistic regressors and R^2 values. * denotes $p \geq 0.05$ (non-significance). Intercept coefficients are small values and we omit them for clarity.

	WMT					STS				
Metric	SEM	SYN	LEX	R^2	**Metric**	SEM	SYN	LEX	R^2
XMoverScore	0.37	-0.08*	0.40	0.39	XMoverScore	0.09	-0.12	0.46	0.24
LASER	0.29	-0.03*	0.38	0.30	LASER	0.59	0.02*	0.27	0.51
mUSE	0.34	0.01*	0.35	0.33	mUSE	0.72	0.02*	0.12	0.59
LaBSE	0.47	-0.05*	0.17	0.30	LaBSE	0.63	0.08	0.26	0.56
mSBERT	0.42	-0.04*	0.25	0.31	mSBERT	0.87	-0.01*	0.05	0.80

Table 4: Regression results for Chinese-English reference-free metrics.

	WMT						STS				
Metric	SEM	SYN	LEX	CLB	R^2	**Metric**	SEM	SYN	LEX	CLB	R^2
XMoverScore	0.18	0.10	0.18	0.59	0.68	XMoverScore	0.08	0.16	0.36	0.48	0.53
LASER	0.18	-0.02*	0.32	0.34	0.43	LASER	0.56	0.03*	0.28	0.17	0.55
mUSE	0.26	0.00*	0.25	0.30	0.40	mUSE	0.72	0.02*	0.12	0.14	0.61
LaBSE	0.37	-0.05*	0.14	0.30	0.40	LaBSE	0.63	0.09	0.25	0.10	0.58
mSBERT	0.35	-0.04*	0.16	0.31	0.40	mSBERT	0.87	-0.01*	0.05*	0.07	0.81

Table 5: Regression results for Chinese-English reference-free metrics. We add the CLB factor in the regression.

ferent participating MT systems as well as slightly different task definitions, corresponding to an aggregation of different domains. The STS dataset is monolingual and was translated by Google Translate for the cross-lingual scenario. The latter may lower the quality of the data, but WMT data also contains translations.

The WMT scores measure the similarity between the reference and the hypothesis but we compare the source with the hypothesis in the cross-lingual scenario, which reflects a mismatch. Fomicheva et al. (2020a) provide a dataset which gives human DA scores between source and hypothesis. We repeated the experiments with this dataset. The full results are shown in Table 12 in the appendix (omitting the CLB factor). The R^2 scores of 3 out of 5 metrics improve (slightly) compared to the WMT dataset for German-English but all R^2 scores are lower for Chinese-English. This means that mismatched DA scores are apparently not the main reason for our low regression fits. With the new

DA scores, all coefficients for SEM are considerably lower. They are in the range of range 0.06 to 0.14 compared to the maximum of 0.47 for WMT. In contrast, all SYN (0-0.12) scores are higher especially vof German. All MOR (0.16-0.33) and most LEX scores are higher but they are still in a similar range as for the original DA scores.

6 Analysis

	Lex(A,B)	Lex(A,C)	Size
Freitag et al. (2020)	0.49	0.99	1452
PAWS	0.84	0.94	100

Table 6: Averaged lexical overlap and size of the datasets for the adversarial experiments.

In the following, we analyze two observations from our previous experiments in more depth: (i) the sensitivity of metrics to lexical overlap; (ii) the orthogonality of metrics in that they capture different linguistic signals.

PAWS		
Sentence A	Sentence B	Sentence C
Later in 2014 , Dassault Systèmes was bought by Quintiq.	Dassault Systèmes was bought in 2014 by Quintiq.	In 2014 , Quintiq was bought by Dassault Systèmes.
They are high , built of concrete faced with small blocks of stone.	They are high built of concrete with small stone blocks.	They are small , built of concrete with high stone blocks.

Freitag et al. (2020) translated		
Sentence A	Sentence B	Sentence C
Shark injures 13-year-old on lobster dive in California	A 13-year-old is injured by a shark while diving for lobsters in California	Shark injures 13-year-old on dive lobster in California
Kovacic did a quick give-and-go at midfield.	Kovacic managed a quick one-two in midfield.	Kovacic did a quick midfield at give-and-go.

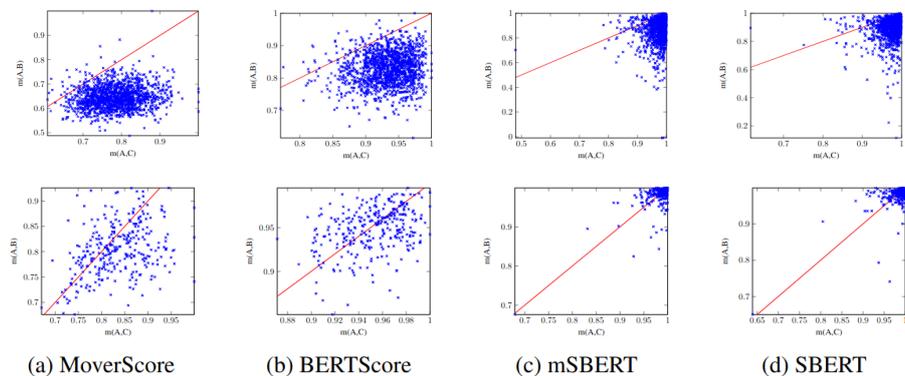Table 7: Selected sentences for the adversarial experiments.

(a) MoverScore (b) BERTScore (c) mSBERT (d) SBERT

Figure 3: Distribution of $\mathfrak{m}(A, B)$ vs. $\mathfrak{m}(A, C)$. Top: Dataset from Freitag et al. (2020). Bottom: PAWS dataset.

6.1 Adversarial experiments

According to our results, all metrics rely on lexical overlap, which indicates that they may not be robust to adversarial examples. We check this by an additional experiment, for reference-based metrics, in which we query their pairwise preferences over three sentences: Sentence A is the anchor sentence; sentence B is a paraphrase of sentence A with little lexical overlap; sentence C is a non-paraphrase with high lexical overlap. A good metric $\mathfrak{m}$ would have $\mathfrak{m}(A, B) > \mathfrak{m}(A, C)$ but the high lexical overlap between A and C makes this task difficult.

Freitag et al. Sentences A are taken as source sentences from WMT19. Freitag et al. (2020) provided alternative references for WMT19; they instructed human professional translators to paraphrase the references as much as possible in terms of lexical choice and sentence structure but keep the same semantics. We take these as sentences B. We produce sentences C from sentences A: for each sentence A, we detect the nouns within the sentence using the NLTK POS tagger, and then we randomly permute them to produce a sentence

C. Since Freitag et al. (2020) provided sentences in German, we translate all sentences into English using Google translate. We note that by inspection the translations are generally of high quality and satisfy our constraints of inducing paraphrases with low lexical overlap and non-paraphrases with high lexical overlap. Examples are shown in Table 7 and statistics in Table 6.

PAWS We complement the analysis with the native English PAWS dataset (Zhang et al., 2019) which consists of paraphrase and non-paraphrase pairs that have high lexical overlap. For each sentence in the dataset, there are multiple paraphrases and non-paraphrases. For a given sentence A, we use the paraphrase with the smallest amount of lexical overlap as sentence B and sentence C is the non-paraphrase with the highest amount of lexical overlap with A. We note that PAWS is more problematic as even the paraphrases B with minimum amount of lexical overlap do have considerable lexical overlap. Therefore, we select the 100 sentence pairs with the smallest amount of lexical overlap between sentence A and B. Table 6 shows the lex-

ical overlap and the size of the datasets. Indeed, for PAWS, sentences B have only a little less lexical overlap with A than sentences C, while the dataset of Freitag et al. (2020) has a much clearer separation between B and C in this respect.

Figure 3 shows the distribution of $\mathfrak{m}(A, B)$ and $\mathfrak{m}(A, C)$ for selected metrics $\mathfrak{m}$. Overall, the adversarial results on translated and non-translated datasets point in a similar direction. We see that metrics clearly prefer the high lexical overlap sentences which are non-paraphrases (sentences C) in the translated dataset of Freitag et al. (2020). For non-translated PAWS, metrics are at least to some degree indifferent, but tend to prefer B on average, with MoverScore and BERTScore having higher preference for C than mSBERT and SBERT, which confirms our linear regression results.

Overall, these experiment show that metrics are indeed not robust to lexical adversarial examples.

6.2 Ensemble of Models

Our experiments in Section 4 show that the different metrics use different information signals, even when they use the same underlying BERT representations. For example, BERTScore relies more on lexical overlap and mSBERT relies more on semantics; BERTScore and MoverScore both capture syntax, while the other metrics are less sensitive to it. This means that the metrics are to some degree orthogonal. Thus, we suspect that a combination of metrics yields a considerably better metric. We check this hypothesis through an extra experiment. We combine especially BERTScore and MoverScore with SBERT and mSBERT based metrics.

We evaluate the metrics on segment-level. In segment-level evaluation, each sentence pair gets a score from $\mathfrak{m}$. The Pearson correlation coefficient is then calculated between these scores and the human judgement, disregarding the systems which generated the translations. To combine metrics, we simply average their scores. In the evaluation, we use the best performance of the single metrics as baseline and compare it to our combined metrics.

Table 13 (appendix) shows the improvements for different language pairs. In the reference-free case, the two best ensembles combine XMover-Score with mSBERT and LaBSE—the latter two rely less on lexical overlap than the first—leading to big improvements of 11-13% over the best individual metric. Combining metrics relying on similar factors shows less improvement, and often even leads to worse results. In the reference-based case, we combine BERTScore with mSBERT and observe an improvement of 8%, more than for any other combination we tested. These results show that combing metrics relying on different factors can largely improve their performance.[1]

7 Conclusions

We disentangled BERT-based evaluation metrics along four linguistic factors: semantics, syntax, morphology, and lexical overlap. The results indicate that (i) the different metrics capture these different aspects to different degrees but (ii) they all rely on semantics and lexical overlap. The first observation indicates that combining metrics may be helpful, which we confirmed: simple parameter-free averaging of hetereogenous metric scores can improve correlations with humans by up to more than 13% in our experiments. The second observation shows that these metrics may be prone to adversarial fooling, just like BLEU and ROUGE, which we confirmed in an additional experiment in which we queried metric preferences over paraphrases with little lexical overlap and non-paraphrases with high lexical overlap. Future metrics should especially take this last aspect into account, and improve their robustness to adversarial conditions.

There is much scope for future research, e.g., in developing better global explanations for reference-free metrics (as we cannot yet well explain these metrics), better linguistic factors (e.g., a clearer conceptualization of morphological similarity of two sentences) and in developing local explainability techniques for evaluation metrics (Fomicheva et al., 2021a,b).[2]

---

[1] In work independent from us, Song et al. (2021) also ensemble BERT based reference-free evaluation metrics.

[2] Code and data for our experiments are available from https://github.com/SteffenEger/global-explainability-metrics.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond.

Philip Bille. 2005. A survey on tree edit distance and related problems. *Theoretical computer science*, 337(1-3):217–239.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. 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, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2020. How to probe sentence embeddings in low-resource languages: On structural design choices for probing task evaluation. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 108–118, Online. Association for Computational Linguistics.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. 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 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021a. The eval4nlp shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2021b. Translation error detection as rationale extraction.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2020a. MLQE-PE: A multilingual quality estimation and post-editing dataset. *arXiv preprint arXiv:2010.04480*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020b. BERGAMOT-LATTE submissions for the WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1010–1017, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Zorik Gekhman, Roee Aharoni, Genady Beryozkin, Markus Freitag, and Wolfgang Macherey. 2020. KoBE: Knowledge-based machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3200–3207, Online. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Yannik Keller, Jan Mackensen, and Steffen Eger. 2021. BERT-defense: A probabilistic model based on BERT to combat cognitively inspired orthographic adversarial attacks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1616–1629, Online. 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. PMLR.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. 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.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest at WMT2020: Sentence-level direct assessment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1049–1055, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. SentSim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156, Online. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Bin Wang and C. C. Jay Kuo. 2020. Sbert-wk: A sentence embedding method by dissecting bert-based word models.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual universal sentence encoder for semantic retrieval.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262.

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*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

8 Appendix

The following tables contain remaining experimental results.

Dataset	English	Chinese	German
WMT	8595	560	1620
STSB	5749	5749	5749

Table 8: Number of sentence pairs in each dataset and language pair.

WMT		
Sentence 1	**Sentence 2**	**SEM**
Why is it such a difference?	Why such a difference?	0.94
There are coincidences.	Zufälle gibt es.	0.57

STS		
Some men are fighting.	Two men are fighting.	4.25
A woman is writing.	Eine Frau schwimmt.	0.1

Table 9: Example of WMT and STS datasets.

	WMT						STS				
Metric	**SEM**	**SYN**	**LEX**	**MOR**	R^2	**Metric**	**SEM**	**SYN**	**LEX**	**MOR**	R^2
XMoverScore	0.37	-0.07*	0.33	-0.01*	0.31	XMoverScore	0.08	-0.02*	0.51	0.03	0.30
LASER	0.21	-0.08	0.31	0.31	0.29	LASER	0.54	0.00*	0.28	0.15	0.54
mUSE	0.19	-0.00*	0.40	0.08	0.25	mUSE	0.71	0.00*	0.09	0.07	0.6
LaBSE	0.30	-0.03*	0.26	0.07	0.21	LaBSE	0.60	0.05	0.29	0.04	0.60
mSBERT	0.39	-0.06*	0.22	0.05*	0.26	mSBERT	0.89	-0.02	0.05	0.02	0.84

Table 10: Regression results for German-English reference-free metrics. Coefficient values for linguistic regressors and R^2 values. * denotes $p \geq 0.05$ (non-significance).

	WMT							STS					
Metric	**SEM**	**SYN**	**LEX**	**MOR**	**CLB**	R^2	**Metric**	**SEM**	**SYN**	**LEX**	**MOR**	**CLB**	R^2
XMoverScore	0.14	0.08	0.16	-0.03*	0.63	0.61	XMoverScore	0.09	0.25	0.43	-0.09	0.40	0.54
LASER	0.11	-0.03*	0.28	0.19	0.40	0.46	LASER	0.51	0.01*	0.32	0.12	0.11	0.58
mUSE	0.19	0.02*	0.28	0.03*	0.35	0.34	mUSE	0.70	0.02*	0.12	0.05*	0.11	0.61
LaBSE	0.29	0.00*	0.14	0.00*	0.45	0.38	LaBSE	0.59	0.06	0.30	0.02*	0.12	0.62
mSBERT	0.36	-0.02*	0.09	0.02*	0.36	0.36	mSBERT	0.88	-0.02*	0.05*	0.02*	0.03*	0.85

Table 11: Regression results for German-English reference-free metrics. Coefficient values for linguistic regressors and R^2 values. * denotes $p \geq 0.05$ (non-significance). We add the CLB factor in the regression.

	German-English						Chinese-English			
Metric	**SEM**	**SYN**	**LEX**	**MOR**	R^2	**Metric**	**SEM**	**SYN**	**LEX**	R^2
XMoverScore	0.16	0.05*	0.33	0.16	0.20	XMoverScore	0.14	0.04*	0.35	0.15
LASER	0.12	0.08	0.30	0.46	0.41	LASER	0.06*	0.03*	0.35	0.14
mUSE	0.12	0.12	0.25	0.43	0.35	mUSE	0.12	0.05*	0.42	0.21
LaBSE	0.10	0.08	0.23	0.51	0.41	LaBSE	0.16	0.04*	0.43	0.23
mSBERT	0.14	0.06*	0.21	0.33	0.22	mSBERT	0.24	0.00*	0.31	0.17

Table 12: Regression results for reference-free metrics on the Fomicheva et al. (2020a) dataset which contains DA scores comparing sources and hypotheses (rather than references and hypotheses). Coefficient values for linguistic regressors and R^2 values. * denotes $p \geq 0.05$ (non-significance).

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Avg
Reference-based								
BERTScore + mSBERT	8%	15%	7%	7%	11%	2%	4%	8%
SBERT + MoverScore	3%	10%	1%	11%	6%	-1%	1%	4%
SBERT + LASER	6%	6%	0%	2%	5%	3%	10%	5%
SBERT + mUSE	5%	10%	4%	2%	8%	4%	2%	5%
Reference-free								
XMoverScore + mSBERT	9%	19%	11%	12%	18%	13%	12%	13%
XMoverScore + LaBSE	8%	14%	5%	10%	17%	9%	17%	11%
mSBERT + LASER	8%	9%	-2%	1%	3%	0%	12%	4%
LaBSE + LASER	7%	15%	-4%	2%	1%	-2%	4%	3%
XMoverScore + LASER	0%	7%	-5%	2%	4%	1%	13%	3%
mSBERT + LaBSE	3%	7%	-11%	-6%	-7%	-6%	2%	-3%

Table 13: Performance gains from the ensemble metrics over single best metrics.