USCORE: An Effective Approach to Fully Unsupervised Evaluation Metrics for Machine Translation

Jonas Belouadi
jonas.belouadi@uni-bielefeld.de

Steffen Eger
steffen.eger@uni-bielefeld.de

Natural Language Learning Group (NLLG)
Faculty of Technology, Bielefeld University
nl2g.github.io

Abstract

The vast majority of evaluation metrics for machine translation are supervised, i.e., (i) are trained on human scores, (ii) assume the existence of reference translations, or (iii) leverage parallel data. This hinders their applicability to cases where such supervision signals are not available. In this work, we develop *fully unsupervised* evaluation metrics. To do so, we leverage similarities and synergies between evaluation metric induction, parallel corpus mining, and MT systems. In particular, we use an unsupervised evaluation metric to mine pseudo-parallel data, which we use to remap deficient underlying vector spaces (iteratively) and to induce an unsupervised MT system, which then provides pseudo-references as an additional component in the metric. Finally, we also induce unsupervised multilingual sentence embeddings from pseudo-parallel data. We show that our fully unsupervised metrics are effective, i.e., they beat supervised competitors on four out of five evaluation datasets. We make our code publicly available.[1]

1 Introduction

Evaluation metrics are essential for judging progress in natural language generation (NLG) tasks such as machine translation (MT) and summarization, as they identify the state-of-the-art in a key NLP technology. Despite their wide dissemination, classical lexical overlap evaluation metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) have difficulties judging the quality of modern NLG systems (Mathur et al., 2020a; Marie et al., 2021), necessitating novel metrics that correlate better with humans. Lately, this has been a very active research area (Zhang et al., 2020; Zhao et al., 2019, 2022; Colombo et al., 2021; Yuan et al., 2021).[2]

Recently, more and more *supervised* metrics are being proposed. E.g., BLEURT (Sellam et al., 2020) trains on human annotated datasets ranging from 5k-150k pairs, the COMET (Rei et al., 2020) models regress on 12k-370k data points and UNITE (Wan et al., 2022), before fine-tuning on the same data as COMET, pre-trains on 5m-10m parallel sentences. Of course, training on larger amounts of data leads to better metrics (measured on in-domain data), but also increases the risk of learning biases from the data (Poliak et al., 2018)—and limits the applicability to domains and language pairs where supervision is available. Here, we go the opposite route and try to minimize the amount of supervision as much as possible.

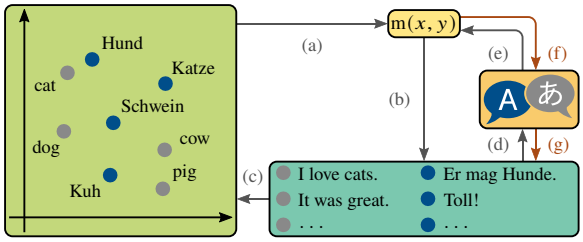


Figure 1: Relationship between metrics m, vector spaces, parallel data, and MT systems: Metrics build on (potentially deficient) multilingual vector spaces (a), and can be used to mine (pseudo-)parallel sentences (b), which in turn can be used to improve deficient vector spaces (c). (Pseudo-)parallel data can also be used to train MT systems (d), which can generate pseudo-references (e). Conversely, metrics can also be optimization criteria for MT systems, which in turn can generate additional pseudo-parallel data through translation (f & g; not explored in this work).

We classify existing metrics making use of different types of supervision as follows (cf. Table 1). **TYPE-1** metrics are *trained* on human assessments such as Direct-Assessment (DA) or Post-Editing (PE) scores, and compare system outputs to either human references (reference-based; Sellam et al., 2020; Rei et al., 2020) or directly to source texts (reference-free; Ranasinghe et al., 2021). **TYPE-2** metrics, by comparison, do not use human assessments for training but still require

[1] github.com/potamides/unsupervised-metrics
[2] Of course, the search for high quality metrics dates back at least to the invention of BLEU and its predecessors.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 358–374
May 2-6, 2023

	Training	References	Parallel Data
TYPE-1	✓	(✓)	(✓)
TYPE-2	✗	✓	(✓)
TYPE-3	✗	✗	✓
Unsupervised	✗	✗	✗

Table 1: Different types of supervision used by TYPE-1/2/3 metrics compared to unsupervised metrics. Checkmarks surrounded by parentheses denote optional supervision signals.

human references, i.e., are untrained and reference-based (Yuan et al., 2021; Zhao et al., 2019; Zhang et al., 2020). Finally, **TYPE-3** metrics are untrained (unlike TYPE-1) and reference-free (unlike TYPE-2), i.e., do not use supervision as in TYPE-1 or 2. However, to work well, they still rely on parallel data (Zhao et al., 2020; Song et al., 2021), which is considered a form of supervision, e.g., in the MT community (Artetxe et al., 2018; Lample et al., 2018).

In contrast, we aim for *fully unsupervised evaluation metrics* (for MT) that do not use any form of supervision (cf. Table 1). In addition, subject to the constraint that no supervision is allowed, our metrics should be of *maximally high quality*, i.e., correlation with human assessments. We have two use cases in mind: (a) Such *sample efficiency*[3] is a prerequisite for the wide applicability of the metrics. This is especially important when we want to overcome the current English-centricity (Anastasopoulos and Neubig, 2020) of MT systems and evaluation metrics and also cover low-resource *languages* like Nepali or Sinhala (Fomicheva et al., 2021) and low-resource *pairs* like Yoruba-German.[4] (b) Our fully unsupervised evaluation metrics should be considered *strong lower bounds* for any future work that uses (mild) forms of supervision for metric induction, i.e., we want to push the lower bounds for newly developed TYPE-K metrics.

To achieve our goals, we employ *self-learning* (He et al., 2020; Wei et al., 2021) and in particular, we leverage the following dualities to make our metrics maximally effective, cf. Figure 1: (1) Evaluation metrics and NLG systems are closely related; e.g., a metric can be an optimization criterion for an NLG system (Böhm et al., 2019), and a system can conversely generate *pseudo references* (a.o.) from which to improve a metric. (2) Evaluation metrics and parallel corpus mining (Artetxe and Schwenk, 2019) are closely related; e.g., a metric can be used to mine parallel data, which in turn can be used to improve the metric (Zhao et al., 2020), e.g., by remapping deficient embedding spaces.

Our contributions are: **(i)** We show that effective unsupervised evaluation metrics can be obtained by exploiting relationships with parallel corpus mining approaches and MT system induction; **(ii)** to do so, we explore ways to (a) make parallel corpus mining *efficient* (e.g., overcome cubic runtime complexity) and (b) induce unsupervised multilingual sentence embeddings from pseudo-parallel data; **(iii)** we show that pseudo-parallel data can rectify deficient vector spaces such as MBERT; **(iv)** we show that our metrics beat three state-of-the-art supervised metrics on four of five datasets we evaluate on.

2 Background

We take inspiration from three recent supervised (reference-free; TYPE-3) metrics: XMoverScore (Zhao et al., 2020), DistilScore (Reimers and Gurevych, 2020), and SentSim (Song et al., 2021). Below, we review key aspects of them, and show where supervision plays a role.

2.1 XMoverScore

Central to XMoverScore is the use of Word Mover's Distance (WMD) as a similarity between two sentences (Zhao et al., 2020). WMD and further enhancements are discussed below.

WMD WMD is a distance function that compares sentences at the token level (Kusner et al., 2015), by leveraging word embeddings which in XMoverScore's case come from MBERT (Devlin et al., 2019). From a source sentence x and an MT hypothesis y, WMD constructs a *distance matrix* $\mathbf{C} \in \mathbb{R}^{|x|,|y|}$, where $\mathbf{C}_{ij}$ is the distance between two word embeddings, $\mathbf{C}_{ij} = \| \mathrm{E}(x_i) - \mathrm{E}(y_j) \|$; x_i, y_j index respective words in x, y. WMD uses this distance matrix to compute the similarity of the two sentences. This can be defined as the linear programming problem

$$\mathrm{WMD}(x, y) = \min_{\mathbf{F}} \sum_{i=1}^{|x|} \sum_{j=1}^{|y|} \mathbf{F}_{ij} \mathbf{C}_{ij}, \qquad (1)$$

[3] We use the term sample efficiency in a generalized sense to denote the amount of supervision required.

[4] Neither Yoruba (a language spoken in Nigeria) nor German are classical low-resource languages. For German, this is clear and Yoruba is even included in MBERT, i.e., belongs to the languages with 100+ largest Wikipedias. Nonetheless, from own experience, we find it inherently difficult to obtain high-quality annotations for the language pair, as a result of few competent parallel speakers as well as technical difficulties (e.g., lack of adequate compute infrastructure in Nigeria).

where $\mathbf{F} \in \mathbb{R}^{|x|,|y|}$ is an *alignment matrix* with $\mathbf{F}_{ij}$ denoting how much of word x_i travels to word y_j. Additional constraints prevent it from becoming a zero matrix.

Vector space remapping Zhao et al. (2020), akin to similar earlier and subsequent work (Cao et al., 2020; Schuster et al., 2019), argue that the monolingual subspaces of MBERT are not well aligned. As a remedy, they investigate linear projection methods which post-hoc improve cross-lingual alignments. We refer to this approach as vector space *remapping*. XMOVERSCORE explores two different remapping approaches, CLP and UMD. They both leverage parallel data on sentence-level from which they extract word-level alignments using FAST-ALIGN, which are then used for remapping. We give more details in the Appendix A.

Language Model XMOVERSCORE linearly combines WMD with the perplexity of a GPT-2 language model (Radford et al., 2019). Allegedly this penalizes ungrammatical translations. This updates XMOVERSCORES scoring function to

$$\mathfrak{m}(x, y) = w_{\text{xlng}}\,\text{WMD}(x, y) + w_{\text{lm}}\,\text{LM}(y). \quad (2)$$

Here, w_{xlng}, w_{lm} are weights for the cross-lingual WMD and LM components of XMOVERSCORE.

2.2 DISTILSCORE

Reimers and Gurevych (2020) show that the cosine between multilingual sentence embeddings captures semantic similarity and can be used to assess cross-lingual semantic textual similarity. Their approach to inducing embedding models is based on multilingual knowledge distillation. We refer to this metric as DISTILSCORE. Their approach requires supervision at multiple levels. First, parallel sentences are needed to induce multilingual models, and second, NLI and STS corpora are required to induce teacher embeddings in the source language.

2.3 SENTSIM

A key difference between XMOVERSCORE and DISTILSCORE is that one approach is based on word- and the other on sentence embeddings. Song et al. (2021) and Kaster et al. (2021) show that combining approaches based on word-level and sentence-level representations can substantially improve metrics. The metric of Song et al. (2021), which is called SENTSIM, combines supervised DISTILSCORE with one of two word embedding-based metrics. The first one is quite similar to XMOVERSCORE, as it is also based on WMD. The other one is a multilingual variant of BERTSCORE (Zhang et al., 2020).

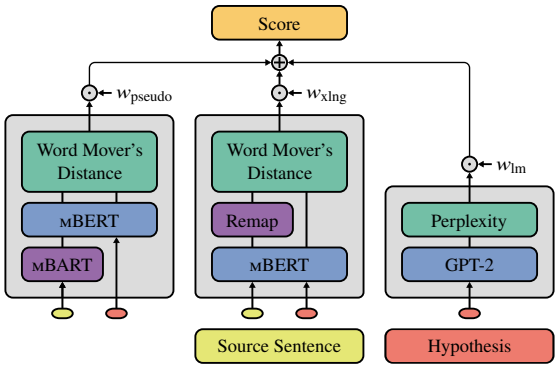


Figure 2: $\text{USCORE}_{\text{WRD}}$ with pseudo references (left), unsupervised remapping (middle) and a LM (right).

3 Methods

In this section, we introduce our fully unsupervised metric USCORE. USCORE builds upon the existing metrics XMOVERSCORE, DISTILSCORE, and SENTSIM, but eliminates all supervision signals and instead leverages the dualities shown in Figure 1. In particular, we mine pseudo-parallel data from unsupervised metrics, which we use (iteratively) to (a) rectify deficient vector spaces (for XMOVERSCORE) and to (b) train unsupervised MT systems which can generate pseudo-references (as pseudo-references are in the same language as the hypothesis, this eliminates problems of cross-lingual deficiency). Furthermore, we use pseudo-parallel data to (c) induce an unsupervised sentence embedding model analogous to DISTILSCORE, which we can then (d) integrate with the unsupervised word based model analogous to SENTSIM. We now give details.

3.1 $\text{USCORE}_{\text{WRD}}$

XMOVERSCORE uses sentence-parallel data to extract word pairs for vector space remapping. (i) We replace this parallel data with pseudo-parallel data.[5] (ii) In addition, we use pseudo-references to address the issue of deficient vector spaces. We now give details on (i) and (ii) below.

Efficient WMD Pseudo-Parallel Data Mining Metrics such as XMOVERSCORE could in principle be used for pseudo-parallel corpus mining since

[5]To extract the word pairs from the sentence-parallel data, XMOVERSCORE uses FAST-ALIGN (Dyer et al., 2013), but since this depends directly on how well sentences are aligned, we first replace it with unsupervised AWESOME-ALIGN (Dou and Neubig, 2021) which only relies on pre-trained language models.

they can compare arbitrary sentences. However, when WMD-based metrics are scaled to corpus mining, algorithmic efficiency problems arise: (a) the computational complexity of WMD scales cubically with sentence length (Kusner et al., 2015); (b) to compare m source to m target sentences, m^2 WMD invocations are necessary, which quickly becomes intractable. Thus, we explore ways to improve the performance of WMD to mine efficiently. In particular, Kusner et al. (2015) define a linear approximation of WMD called *word centroid distance* (WCD) and a mining algorithm that first sorts all target samples according to their WCD to a given query and computes exact WMD for the k nearest neighbors. We use this algorithm for efficient WMD-based pseudo-parallel data mining.

In our work, we apply this approach *iteratively* (cf. Figure 1): we start out with an initial WMD metric (based on mBERT), obtain *sentence-level* pseudo-parallel data with it via the efficient approximation algorithm described, and obtain a dictionary of word pairs from unsupervised awesome-align from the sentence pseudo-parallel data. We use the pseudo-parallel word pairs with UMD and CLP to remap mBERT. From this, we obtain a better WMD metric; then we iterate.

Pseudo References Apart from remapping, pseudo-parallel data could be used to overcome problems of deficient vector spaces in other ways. Specifically, we want to mine enough pseudo-parallel data to train an unsupervised MT system to translate source sentences into the target language to create *pseudo references* (Albrecht and Hwa, 2007; Gao et al., 2020; Fomicheva et al., 2020b). This would allow for a comparison with the hypothesis in the target language, similar to reference-based metrics, circumventing alignment problems in multilingual embeddings. This approach updates $\text{USCORE}_{\text{WRD}}$ to

$$\mathfrak{m}(x, y, y') = w_{\text{xlng}}\,\text{WMD}^{(n)}(x, y) + w_{\text{lm}}\,\text{LM}(y) + w_{\text{pseudo}}\,\text{WMD}(y, y'), \tag{3}$$

where n denotes the iterations of remapping, and w_{pseudo} is a new weight to control the influence of the pseudo reference y'. All components of $\text{USCORE}_{\text{WRD}}$ are illustrated in Figure 2.

3.2 $\text{USCORE}_{\text{SNT}}$

Besides a word-based metric, we use pseudo-parallel data to induce an unsupervised *sentence level* metric, $\text{USCORE}_{\text{SNT}} = \cos(x, y)$, based on the cosine similarity between sentence embeddings. One could, similarly to DistilScore, perform knowledge distillation but since our initial experiments showed that this doesn't work well with pseudo-parallel data, we chose another approach.

Contrastive Learning We explore contrastive learning for unsupervised *multilingual* sentence embedding induction, which has recently been successfully used to train unsupervised *monolingual* sentence embeddings (Gao et al., 2021). In our context, the basic idea is to pull semantically close sentences together and to push distant sentences apart in the embedding space. Let x_i and y_i be the embeddings of two sentences that are semantically related and N an arbitrary batch size. The training objective for this pair can be formulated as

$$L_i = -\log \frac{\exp\left(\frac{\cos(x_i, y_i)}{\tau}\right)}{\sum_{j=1, j\neq i}^{N} \exp\left(\frac{\cos(x_i, y_j)}{\tau}\right)}, \tag{4}$$

where τ is a temperature hyperparameter that can be used to either amplify or dampen the assessed distances. For each sentence x_i, all remaining sentences $y_{j\neq i}$ in the current batch should be pushed apart in the embedding space. For positive sentences that should be pulled together, we again use pseudo-parallel sentence pairs. Since noisy data is beneficial for contrastive learning (Gao et al., 2021), we expect this paradigm to work well with pseudo-parallel data. We use pooled XLM-R embeddings as sentence representations, and, as with unsupervised remapping, we experiment with multiple iterations of successive mining and sentence embedding induction operations.

Ratio Margin Pseudo-Parallel Data Mining As $\text{USCORE}_{\text{SNT}}$ is based on sentence embeddings, we cannot use the WMD-based mining algorithm to obtain pseudo-parallel sentences since it requires access to word-level representations. An alternative would be to just use cosine similarity for mining, but that approach is susceptible to noise in the data (Artetxe and Schwenk, 2019). Instead, we follow Artetxe and Schwenk (2019) and use a ratio margin function defined as

$$\text{margin}(x, y) = \frac{\cos(x, y)}{\sum\limits_{z\in N_x} \frac{\cos(x,z)}{2k} + \sum\limits_{z\in N_y} \frac{\cos(y,z)}{2k}}, \tag{5}$$

where N_x and N_y are the k nearest neighbors of sentence embeddings x and y in the respective

language. Informally, this ratio margin function divides the cosine similarities of the nearest neighbor by the average similarities of the neighborhood.

3.3 USCORE$_{\text{WRD} \oplus \text{SNT}}$

Inspired by SENTSIM, which combines word and sentence embeddings, we similarly ensemble USCORE$_{\text{WRD}}$ and USCORE$_{\text{SNT}}$. We refer to this final metric as USCORE = USCORE$_{\text{WRD} \oplus \text{SNT}}$ with two new weights w_{wrd} and $w_{\text{snt}} = 1 - w_{\text{wrd}}$:

$$\begin{aligned} \text{USCORE}(x, y) = w_{\text{wrd}}\, \text{USCORE}_{\text{WRD}}(x, y) \\ + w_{\text{snt}}\, \text{USCORE}_{\text{SNT}}(x, y). \end{aligned} \quad (6)$$

4 Experiments

In this section, we evaluate all USCORE variants at the segment level[6] and compare them to TYPE-1/2/3 upper bounds. We detail additional hyperparameters in Appendix D.

4.1 Datasets

We use various datasets to assess the performance of our metrics on **MT evaluation**, i.e., computing the correlation with human assessments using Pearson's r correlation, and **parallel sentence matching**, a standard evaluation measure in the corpus mining field where a set of shuffled parallel sentences is searched to recover correct translation pairs (Guo et al., 2018; Kvapilíková et al., 2020). For this we report Precision at N (P@N).

MT evaluation In **WMT-16** and **WMT-17,** each language pair consists of tuples of source sentences, hypotheses and references. Each tuple was annotated with a direct assessment (DA) score, which quantifies the adequacy of the hypothesis given the reference translation. Following Zhao et al. (2020) and Song et al. (2021), we use these DA scores to assess the adequacy of the hypothesis given the source. **MLQE-PE** has been used in the WMT 2020 Shared Task on Quality Estimation (Specia et al., 2020), and only provides source sentences and hypotheses for its language pairs, with no references. Each source sentence and hypothesis pair was annotated with cross-lingual direct assessment (CLDA) scores. In terms of annotation, **Eval4NLP** is very similar to MLQE-PE but focuses on non-English-centric language directions, especially de-zh and ru-de. **WMT-MQM** uses fine-grained error annotations from the Multidimensional Quality Metrics (MQM) framework (Freitag et al., 2021a) for adequacy assessments. Like MLQE-PE and Eval4NLP, WMT-MQM also assigns scores based on source sentences and hypotheses. Additional statistics can be found in the appendix in Table 8. Using ISO 639-1 codes, our datasets cover the language pairs de-zh, ru-de, en-ru, en-zh, cs-en, de-en, en-de, et-en, fi-en, lv-en, ne-en, ro-en, ru-en, si-en, tr-en, zh-en.

Parallel sentence matching To evaluate on parallel sentence matching, we use the **News Commentary**[7] dataset. It consists of parallel sentences crawled from economic and political data.

4.2 Fine-grained analysis on de-en

To gain an understanding of the properties of the iterative techniques and the influence of individual parameters / components, we conduct a fine-grained analysis on the de-en language direction of WMT-16 (for MT evaluation) and News Commentary v15 (for parallel sentence matching). We list examples of pseudo-parallel data used during training in the appendix in Table 3. The mined sentences are often semantically similar, but contain factuality errors (e.g., have wrong places or numbers in hypotheses).

Vector space remapping We explore if remapping works with pseudo-parallel data. We use News Crawl for mining. We randomly extract 40k monolingual sentences per language, and select the top 5% sentence pairs with the highest metric scores for remapping. This gives us the same number of sentences (2k pairs) as were used for remapping XMOVERSCORE.

The results for UMD and CLP-based remapping on de-en can be seen in Figure 3 (top). The figure contains two graphs, one for correlation with human judgments and one for precision on parallel sentence matching. Each graph illustrates model performance before remapping (Iteration 0) and after remapping one to five times. After remapping once, both UMD and CLP improve substantially in Pearson's r correlation. The improvement of CLP, however, is noticeably larger. For subsequent iterations, UMD seems to continue to improve slightly, but the correlations of CLP seem to drop. This can be explained by the results for precision where the P@1 of CLP drops each iteration, meaning the remapping capabilities of the metrics decrease. UMD does not exhibit this problem. Thus, UMD

[6] We do not evaluate at the system level since metrics there often perform very similarly, making it difficult to determine the best metric (Mathur et al., 2020b; Freitag et al., 2021b).

[7] data.statmt.org/news-commentary

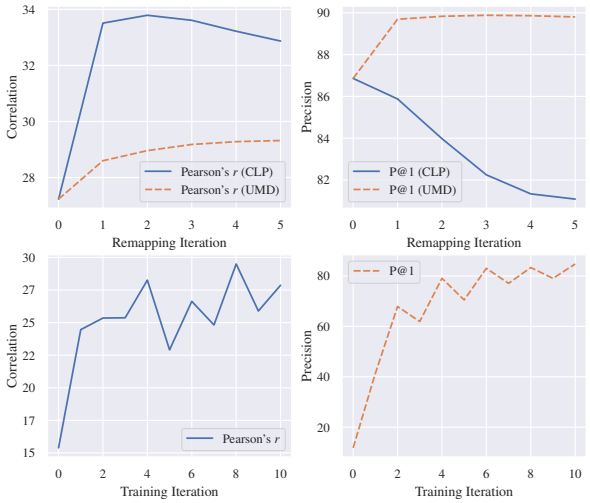


Figure 3: Results for unsupervised vector space remapping (top) and contrastive learning with USCORE$_{\text{SNT}}$ (bottom) for de-en. Pearson's r is computed on WMT-16 (MT evaluation) and P@1 on News Commentary v15 (parallel sentence matching).

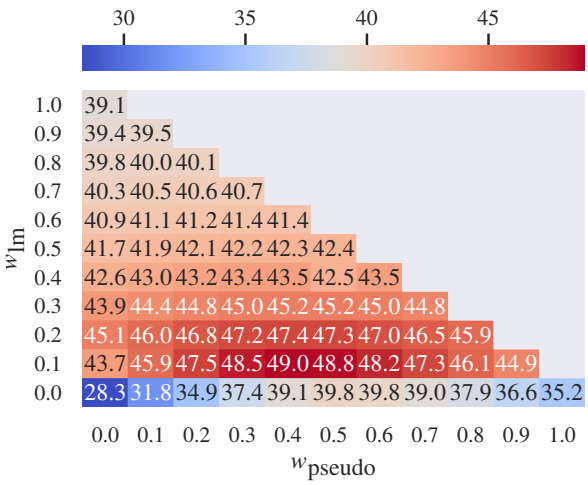


Figure 4: Influence of a language model and an MT system on USCORE$_{\text{WRD}}$, segment-level Pearson's r for different values of w_{pseudo} (weight for pseudo references) and w_{lm} (weight for the language model) on WMT-16 de-en data. Note that the point $w_{\text{pseudo}} = w_{\text{lm}} = 0$ uses only WMD(x, y); see Equation 3. Here, $n = 1$.

could be a more robust choice for metrics that should perform reasonably well on both tasks.

Pseudo References & Language Model Next, we add a language model to the metric and investigate pseudo-parallel corpus mining to train an MT system for pseudo references. Tran et al. (2020) show that fine-tuning MBART using pseudo-parallel data leads to very promising results, so we use MBART for our own experiments as well. Since fine-tuning for MT is a very resource-intensive undertaking requiring many parallel sentence pairs (Barrault et al., 2020), especially compared to our vector space remapping experiments, we need considerably more training data. On average, Tran et al. (2020) use around 200k pseudo-parallel sentence pairs for training. To obtain the same amount with our extraction rate of 5%, we now use a pool of 4m sentences per language for mining. Our results on the de-en data of WMT-16 are reported in Figure 4, which is similar to an ablation study. On the x-axis, we vary the weight $w_{\text{pseudo}} \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$ for USCORE$_{\text{WRD}}$ with pseudo references, and on the y-axis, we explore different weights $w_{\text{lm}} \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$ for the language model. We set $w_{\text{xlng}} = 1 - w_{\text{pseudo}} - w_{\text{lm}}$. The best correlation uses $w_{\text{pseudo}} = 0.4$, $w_{\text{lm}} = 0.1$, and $w_{\text{xlng}} = 0.5$. The improvement when pseudo references and a language model are included is substantial (over only using WMD)—e.g., we improve from 28% correlation with humans ($w_{\text{pseudo}} = w_{\text{lm}} = 0$) to 49% with the best weight combination, an improvement of 75%.

Contrastive Learning For USCORE$_{\text{SNT}}$, we also use 4m monolingual sentences per language for mining but only retain the top 100k sentence pairs, as for the contrastive training objective we additionally have to filter out duplicate sentences. The results of USCORE$_{\text{SNT}}$ are shown in Figure 3 (bottom). The P@1 scores seem to steadily improve every two training iterations. Beginning with the sixth iteration, the precision seems to converge.

4.3 Other Languages: Results & Analysis

We now test our metrics on other languages and datasets. For USCORE$_{\text{SNT}}$, we train its sentence embedding model for six iterations. For USCORE$_{\text{WRD}}$, we remap MBERT once with UMD and make use of a language model and pseudo references obtained from an MT system. Based on Section 4.2, we set $w_{\text{pseudo}} = 0.4$, $w_{\text{lm}} = 0.1$, and $w_{\text{xlng}} = 0.5$. Additionally, based on analogous, unreported experiments, we set $w_{\text{wrd}} = 0.6$ and $w_{\text{snt}} = 0.4$ for ensembling. Since determining the weights for USCORE$_{\text{WRD}}$ and USCORE$_{\text{WRD} \oplus \text{SNT}}$ this way constitutes a form of supervision, we also evaluate weights chosen independently from our conducted experiments. Namely, we also evaluate USCORE^{+} with $w_{\text{wrd}} = w_{\text{snt}} = 0.5$. For w_{lm}, we follow XMOVERSCORE and set it to 0.1 (setting w_{lm} lower makes sense because the LM only addresses the hypoth-

esis without considering the source); accordingly, we set $w_{\text{xlng}} = w_{\text{pseudo}} = 0.45$. Since $w_{\text{lm}} = 0.1$ coincides with our findings in Section 4.2, we also evaluate USCORE^{++} where each component uses entirely uniform weights, i.e., $w_{\text{wrd}} = w_{\text{snt}} = \frac{1}{2}$ and $w_{\text{lm}} = w_{\text{xlng}} = w_{\text{pseudo}} = \frac{1}{3}$.

Correlations with human judgments averaged over language pairs are shown in Table 2 (individual results are in the appendix). We also present the results of the popular TYPE-2 metric BLEU, where possible, and the recent TYPE-1 metrics MONOTRANSQUEST (Ranasinghe et al., 2020b,a) and COMET-QE (Rei et al., 2021). Finally, as more direct competitors, we compare to the TYPE-3 metrics XMOVERSCORE, SENTSIM, and DISTILSCORE. We compute all reported scores ourselves.

Overall, the tuned weights of USCORE perform marginally better than USCORE^{+} on most datasets, but USCORE^{+} is usually a very close second. USCORE^{++} performs worse, however, and only competitively on two of the five datasets. This indicates that the language model should be set to a lower value, a choice that makes intuitively sense.

Expectedly, DISTILSCORE, which uses parallel data, is always better than USCORE$_{\text{SNT}}$, which uses pseudo-parallel data. In contrast, USCORE$_{\text{WRD}}$ is generally on par with XMOVERSCORE, even though XMOVERSCORE uses real parallel data—the difference is that USCORE$_{\text{WRD}}$ also leverages pseudo-references which XMOVERSCORE does not. Indeed, from Figure 4, we observe that the pseudo-references can make an improvement of up to 1–11 points in correlation (comparing 'column' labeled $w_{\text{pseudo}} = 0$ to the columns $w_{\text{pseudo}} > 0$).

Our metrics beat reference-based TYPE-2 BLEU across the board. TYPE-1 metrics, which are fine-tuned on human scores, are generally the best. Intriguingly, the only two language pairs where our metrics are on par with them are the non-English de-zh and ru-de from Eval4NLP. These languages are outside the training scope of the current TYPE-1 metrics and thus test their generalization abilities. For example, on ru-de our best metric outperforms MONOTRANSQUEST by 5 points correlation and COMET-QE by 9 points (Table 7 in the appendix).

USCORE and USCORE^{+} also outperform the TYPE-3 upper bounds on four of five datasets. On WMT-16, WMT-17 and Eval4NLP, they have the best overall results. On WMT-MQM, USCORE$_{\text{WRD}}$ alone is best. The drop in performance for the combined metric is caused by USCORE$_{\text{SNT}}$, which on its own performs very badly. As supervised DISTILSCORE exhibits the same issues, this could be a general problem for sentence embeddings based metrics on this dataset. We identify further reasons in Appendix E.

For MLQE-PE, the SENTSIM metrics perform best on average among TYPE-3 and our metrics—although our reproduced scores for this dataset differ noticeably from the authors' results, due to issues in their original code (Chen et al., 2022). Among our self-learned metrics, the combined variant performs best on average again, but still is 3–5 points below SENTSIM and DISTILSCORE, even though it outperforms both XMOVERSCORE variants by over 6 points. Interestingly, USCORE$_{\text{SNT}}$ works better than USCORE$_{\text{WRD}}$, unlike for the other datasets. Similarly, DISTILSCORE clearly outperforms XMOVERSCORE. This could be because MLQE-PE contains Sinhala, a language MBERT was not trained on. Another explanation is the data collection scheme for ru-en, which uses different sources of parallel sentences, mainly colloquial data and Russian proverbs, which use rather unconventional grammar (Fomicheva et al., 2022). This apparently confuses the language model and MT system which have been trained on data from other domains. When we exclude si-en and ru-en from MLQE-PE, USCORE$_{\text{WRD} \oplus \text{SNT}}$ performs best, with an average Pearson's r of 44.22 for tuned weights and 44.45 for default weights vs. 43.82 for SENTSIM (BERTSCORE). In Appendix C, we show that incorporating real parallel data (in addition to pseudo-parallel data) at an order of magnitude lower than SENTSIM allows us to outperform SENTSIM on MLQE-PE also.

5 Discussion

Throughout, we have presented a mix of results and analysis, which we now summarize and discuss. In Figure 4, we conducted an ablation study on the individual components of USCORE$_{\text{WRD}}$ (on de-en). This showed that all three components (pseudo-references, language model, WMD) matter; by itself, the LM is more important than the pseudo-references which are more important than WMD. However, in combination, the LM is least important. We also showed that pseudo-parallel data can successfully rectify deficient multilingual vector spaces, similar to real parallel data, see Figure 3. We note, however, that pseudo-parallel data may introduce an important bias in our data sampling: namely, it may mine factually incorrect parallel

	Metric	WMT-16	WMT-17	MLQE-PE	Eval4NLP	WMT-MQM
TYPE-1/2 Supervised	MonoTransQuest	64.68	66.47	66.28	42.21	42.23
	COMET-QE	65.76	68.77	49.97	40.00	45.89
	BLEU	47.82	47.01	–	–	19.76
TYPE-3 Supervised	XMoverScore (UMD)	49.96	51.02	33.99	29.60	20.07
	XMoverScore (CLP)	53.10	55.09	31.98	36.93	20.59
	SentSim (BERTScore)	51.86	55.57	**45.36**	26.60	13.71
	SentSim (WMD)	50.66	54.29	44.72	24.44	12.24
	DistilScore	43.79	51.22	43.31	28.90	2.20
Unsupervised	$\text{UScore}_{\text{WRD}}$	53.87	53.52	31.13	36.96	**23.28**
	$\text{UScore}_{\text{SNT}}$	36.06	42.68	37.51	20.39	-0.47
	$\text{UScore}_{\text{WRD} \oplus \text{SNT}}$	**55.78**	**57.55**	40.60	**39.87**	19.35
	$\text{UScore}^{+}_{\text{WRD}}$	53.66	52.97	30.94	36.29	23.15
	$\text{UScore}^{+}_{\text{WRD} \oplus \text{SNT}}$	55.28	57.29	41.41	38.22	17.71
	$\text{UScore}^{++}_{\text{WRD}}$	51.88	51.73	23.76	25.42	19.00
	$\text{UScore}^{++}_{\text{WRD} \oplus \text{SNT}}$	54.99	56.49	34.15	32.05	17.08

Table 2: Segment-level Pearson's r correlations with human judgments, averaged over language directions. The best results are highlighted in bold, while results that are not significantly worse (as determined by a two-sample t-test with 10% significance level) are underlined.

sentences, see Table 3 in the appendix, which may amplify issues of adversarial robustness; see our discussion in Section 8.

We remark that, depending on the annotation scheme, better correlations with human judgments do not necessarily entail better metrics (Freitag et al., 2021a). The datasets in this work were annotated using either DA, CLDA, or MQM scores, with MQM explicitly addressing this problem. Since our metrics are consistent regardless of the annotation scheme, they are unlikely to overfit a particular one.

We finally note that combining word and sentence-level models is meaningful, because they offer complementary views (Song et al., 2021). Kaster et al. (2021) also show that they capture orthogonal linguistic factors to varying degrees. Such complementarity may also stem from different underlying vector spaces, i.e., mBERT vs. XLM-R that we use in sentence- and word-level metrics.

6 Related Work

All metrics in this work presented so far treated the MT model generating the hypotheses as a black-box. There also exists a recent line of work of so-called glass-box metrics, which actively incorporate the MT model under test into the scoring process (Fomicheva et al., 2020a,b). In particular, Fomicheva et al. (2020b) explore whether the MT model under test can be used to generate additional hypotheses (Dreyer and Marcu, 2012). A crucial difference to our metrics is the required availability of the original MT model, which we are agnostic about. The MT models used in Fomicheva et al. (2020b) are all trained on parallel data, which makes their approach a supervised metric in our sense.

Other recent metrics that leverage the relationship between metrics and (MT) systems are Prism (Thompson and Post, 2020) and BARTScore (Yuan et al., 2021). We do not classify them as unsupervised, however, as Prism is trained from scratch on parallel data and BARTScore uses a BART model fine-tuned on labeled summarization or paraphrasing datasets.

There are also multilingual sentence embedding models which are highly relevant in our context. Kvapilíková et al. (2020), for example, fine-tune XLM-R on synthetic data translated with an unsupervised MT system. Similar to our contrastive learning approach, the resulting embedding model is completely unsupervised. Important differences are that our sentence embedding model can be improved iteratively and does not rely on an MT system. We leave a comparison to future work.

Finally, the idea of fully unsupervised text generation systems has originated in the MT community (Artetxe et al., 2018; Lample et al., 2018; Artetxe et al., 2019). Given the similarity of MT systems and evaluation metrics, designing fully unsupervised evaluation metrics is an apparent next step, which we take in this work.

7 Conclusion

In this work, we aimed for sample efficient evaluation metrics that do not use any form of supervision.

In addition, our novel metrics should be maximally effective, i.e., of high quality. To achieve this, we leveraged pseudo-parallel data obtained from fully unsupervised evaluation metrics in three ways: we (i) remapped deficient vector spaces using the pseudo-parallel data, (ii) trained an unsupervised MT system from it (yielding pseudo references), and (iii) induced unsupervised multilingual sentence embeddings. To enable our approach, we also explored *efficient* pseudo-parallel corpus mining algorithms based on our metrics as an orthogonal contribution. Finally, we showed that our approach is effective and can outperform three supervised upper bounds (making use of parallel data) on 4 out of 5 datasets we included in our comparison.

In future work, we want to aim for algorithmic efficiency, include *pseudo source texts* as additional components (using the MT system in backward translation), and address the missing dualities discussed in Figure 1 (i.e., use of metrics as optimization criteria and MT systems to generate additional pseudo-parallel data). Further, our approach has substantial room for improvement given that we selected hyperparameters completely unsupervised or based on one high-resource language pair (de-en). Thus, it will be particularly intriguing to explore *weakly-supervised approaches* which leverage minimal forms of supervision.

8 Limitations

Limitations of our metrics include (1) *algorithmic inefficiency*, (2) *resource inefficiency*, (3) the brittleness of unsupervised MT systems in certain situations, and (4) issues of adversarial robustness.

(1) Some of the components of USCORE$_{\text{WRD}}$ (mainly the MT system) have high computational costs. For example, XMOVERSCORE and SENTSIM (BERTSCORE) take less than 30 seconds to score 1000 hypotheses on an Nvidia V100 GPU. USCORE$_{\text{WRD}}$, on the other hand, takes over 2.5 minutes. This algorithmic inefficiency trades off with our sample efficiency, by which we did not use any supervision signals. In future work, we aim to experiment with efficient MT architectures to reduce computational costs (Kamal Eddine et al., 2022; Grünwald et al., 2022).

(2) Similarly to XMOVERSCORE, MONOTRANSQUEST or SENTSIM, our metrics use high-quality encoders such as MBERT, which are not only memory and inference inefficient but also leverage large monolingual resources. Future work should thus not only investigate using smaller MBERT models but also models that leverage smaller amounts of monolingual resources. Wang et al. (2020), for example, propose a competitive LSTM-based approach that completely forgoes monolingual resources and instead uses small parallel corpora (i.e., a few hundred parallel sentences as a weak supervision signal). Similarly, we give a recipe for improving MBERT for unseen languages using limited amounts of parallel data in Appendix C.

(3) Using unsupervised MT approaches, as we do via pseudo references, may be less effective for truly low-resource languages (Marchisio et al., 2020). However, this remains a very active research field with a constant influx of more powerful solutions (Ranathunga et al., 2022; Sun et al., 2021).

(4) As indicated in Sections 4.2 and 5, our mined pseudo-parallel data tends to contain factual inconsistencies such as "Uruguay was seventh" vs. (a translation of) "Russia was second". As a consequence, our induced metrics may be less robust than existing metrics (Chen and Eger, 2022; Rony et al., 2022). An approach to address this inconsistency would be to retain only high probability aligned words in parallel sentences (recall that we infer word-level parallel data from sentence-level parallel data).

9 Acknowledgments

We thank all reviewers for their valuable feedback, hard work, and time. The last author was supported by DFG grant EG 375/5–1.

References

Joshua Albrecht and Rebecca Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 296–303, Prague, Czech Republic. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2020. Should all cross-lingual embeddings speak English? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8658–8679, Online. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *ACL*.

Loïc Barrault, Magdalena Biesialska, Ondrej Bojar, Marta Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi kiu Lo, Nikola Ljubesic, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt20). In *WMT@EMNLP*.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. 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 3110–3120, Hong Kong, China. Association for Computational Linguistics.

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

Yanran Chen, Jonas Belouadi, and Steffen Eger. 2022. Reproducibility issues for BERT-based evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2965–2989, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanran Chen and Steffen Eger. 2022. Menli: Robust evaluation metrics from natural language inference. *ArXiv*, abs/2208.07316.

Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021. Automatic text evaluation through the lens of Wasserstein barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10466, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. In *AISTATS*.

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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *EACL*.

Markus Dreyer and D. Marcu. 2012. Hyter: Meaning-equivalent semantics for translation evaluation. In *NAACL*.

Sharad Duwal, Amir Manandhar, Saurav Maskey, and Subash Hada. 2019. Efforts in the development of an augmented english-nepali parallel corpus. Technical report, Kathmandu University.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *HLT-NAACL*.

M. Fomicheva, Shuo Sun, Lisa Yankovskaya, F. Blain, Francisco Guzmán, M. Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020a. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marina Fomicheva, Lucia Specia, and Francisco Guzmán. 2020b. Multi-hypothesis machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1218–1232, Online. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *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.

Jens Grünwald, Chris Leiter, and Steffen Eger. 2022. Can we do that simpler? simple, efficient, high-quality evaluation metrics for nlg. *ArXiv*, abs/2209.09593.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Matthew Cer, G. Ábrego, K. Stevens, Noah Constant, Yun-Hsuan Sung, B. Strope, and R. Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *WMT*.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *Proceedings of ICLR*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. 2022. FrugalScore: Learning cheaper, lighter and faster evaluation metrics for automatic text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1305–1318, Dublin, Ireland. Association for Computational Linguistics.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of BERT-based evaluation metrics by disentangling along linguistic factors. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8912–8925, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A. Smith. 2021. Unsupervised bitext mining and translation via self-trained contextual embeddings. *Transactions of the Association for Computational Linguistics*, 8:828–841.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *ICML*.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondrej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *ACL*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. 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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

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. 2020a. 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.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages

180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI Blog*.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020a. Transquest at wmt2020: Sentence-level direct assessment. In *Proceedings of the Fifth Conference on Machine Translation*.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020b. Transquest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2021. An exploratory analysis of multilingual word-level quality estimation with cross-lingual transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 434–440, Online. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2022. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, 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. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP*.

Md Rashad Al Hasan Rony, Liubov Kovriguina, Debanjan Chaudhuri, Ricardo Usbeck, and Jens Lehmann. 2022. RoMe: A robust metric for evaluating natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5645–5657, Dublin, Ireland. Association for Computational Linguistics.

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.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *EACL*.

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

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. Sentsim: Crosslingual semantic evaluation of machine translation. In *NAACL*.

Lucia Specia, F. Blain, M. Fomicheva, E. Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the wmt 2020 shared task on quality estimation. In *WMT@EMNLP*.

Haipeng Sun, Rui Wang, Masao Utiyama, Benjamin Marie, Kehai Chen, Eiichiro Sumita, and Tiejun Zhao. 2021. Unsupervised neural machine translation for similar and distant language pairs: An empirical study. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–17.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *EMNLP*.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. 2021. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*.

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

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

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*.

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 *ACL*.

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.

Wei Zhao, Michael Strube, and Steffen Eger. 2022. DiscoScore: Evaluating text generation with BERT and discourse coherence.

A Vector Space Remapping

Zhao et al. (2020) explore two different remapping approaches for XMoverScore, which are defined as follows:

Procrustes alignment Mikolov et al. (2013) propose to compute a linear transformation matrix $\mathbf{W}$ which can be used to map a vector x of a source word into the target language subspace by computing $\mathbf{W}x$. The transformation can be computed by solving the problem

$$\min_{\mathbf{W}} ||\mathbf{WX} - \mathbf{Y}||_2. \quad (7)$$

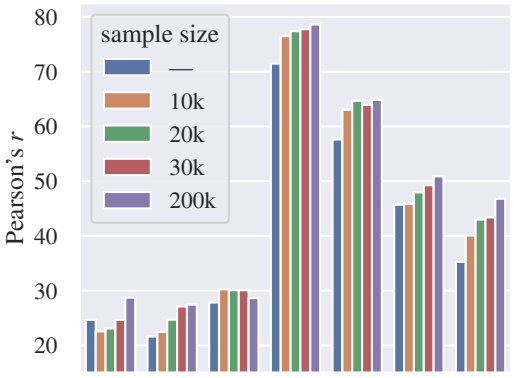
Here $\mathbf{X}, \mathbf{Y}$ are matrices with embeddings of source and target words, respectively, where the tuples $(x_i, y_i) \in \mathbf{X} \times \mathbf{Y}$ come from parallel word pairs. XMoverScore constrains $\mathbf{W}$ to be an orthogonal matrix such that $\mathbf{W}^\intercal\mathbf{W} = \mathbf{I}$, since this can lead to further improvements (Xing et al., 2015). Zhao et al. (2020) call this remapping Linear Cross-Lingual Projection remapping (CLP).

De-biasing The second remapping method of XMoverScore is rooted in the removal of biases from word embeddings. Dev and Phillips (2019) explore a bias attenuation technique called Universal Language Mismatch-Direction (UMD). It involves a bias vector v_B, which is supposed to capture the bias direction. For each word embedding e, an updated word embedding e' is computing by subtracting their projections onto v_B, as in

$$e' = e - (e \cdot v_B)v_B, \quad (8)$$

where $\cdot$ is the dot product. To obtain the bias vector v_B, Dev and Phillips (2019) use a set $\mathcal{E}$ of word pairs that should be de-biased (e.,g. *man* and *woman*). The subtractions of the embeddings of the words in each pair are then stacked to form a matrix $\mathbf{Q}$, and the bias vector v_B is its top-left singular vector. Zhao et al. (2020) use the same approach for XMoverScore, but $\mathcal{E}$ instead consists of parallel word pairs.

Zhao et al. (2020) show that these remapping methods lead to substantial improvements of their XMoverScore metric (on average, up to 10 points in correlation). The required parallel word pairs were extracted from sentences of the EuroParl corpus (Koehn, 2005) using the fast-align (Dyer et al., 2013) word alignment tool. The best results were obtained when remapping on 2k parallel sentences.



	—	10k	20k	30k	200k
Avg	40.60	42.95	44.42	45.17	**46.57**

Figure 5: Pearson's r correlations on MLQE-PE for $\text{UScore}_{\text{WRD} \oplus \text{SNT}}$ when fine-tuning on limited amounts of parallel data. We explore sample sizes of 10k, 20k, 30k, and 200k.

B Filtering

Since large corpora tend to include low-quality data points, we follow Artetxe and Schwenk (2019) and Keung et al. (2021) and apply three simple filtering techniques. We first remove all sentences from each monolingual corpus for which the fastText language identification tool (Joulin et al., 2017) predicts a different language. We then filter all sentences which are shorter than 3 tokens or longer than 30 tokens. As the last step, we discard sentence pairs sharing substantial lexical overlap, which prevents degenerate alignments of, e.g., proper names. We remove all sentence pairs for which the Levenshtein distance detects an overlap of over 50%.

C Fine-Tuning on Parallel Data

To examine whether and by how much we can further improve our metrics using forms of supervision, we experiment with a fine-tuning step on *parallel sentences* and treat self-learning on pseudo-parallel data as pre-training (He et al., 2020). We use the parallel data to fine-tune the contrastive sentence embeddings of $\text{UScore}_{\text{SNT}}$ and the MT system of $\text{UScore}_{\text{WRD}}$, which is responsible for generating pseudo references. Further, we also compute new remapping matrices for $\text{UScore}_{\text{WRD}}$. Since CLP is superior to UMD when parallel data is used (see Section 4.2), we compute these remapping matrices using CLP instead of UMD. To assess how different

amounts of parallel sentences affect performance, we fine-tune our metrics on 10k, 20k, 30k, and 200k parallel sentences. We use WikiMatrix (Schwenk et al., 2021) and the Nepali Translation Parallel Corpus (Duwal et al., 2019) to obtain parallel sentences.

Pearson's r correlations with human judgments for individual and averaged language pairs are shown in Figure 5; we focus on MLPQE-PE, where our metrics performed worst. Overall, introducing parallel data into the training process consistently improves performance for the majority of language directions; more parallel data leads to better results. The relatively biggest improvements are achieved for the si-en language direction, which is in accordance with our discussion above. When fine-tuning with 30k parallel sentences, the performance of our metrics is roughly on par with the SentSim variants (see Table 2). With 200k parallel sentences, our metrics clearly outperform SentSim, which uses millions of parallel sentences and NLI data as supervision signals.

D Hyperparameters

For efficient WMD pseudo-parallel mining, we set $k = 20$ for remapping, and $k = 1$ for training mBART. For ratio-margin-based pseudo-parallel mining, we use $k = 5$. With regard to training $\text{UScore}_{\text{SNT}}$, we follow Gao et al. (2021), and iteratively train XLM-R for one training epoch with a learning rate of 5e-5, a batch size of 256, and a temperature coefficient of $\tau = 0.05$ utilizing the AdamW optimizer (Loshchilov and Hutter, 2019). Fine-tuning of mBART was performed for three epochs with a batch size of four and using the same learning rate of 5e-5 as well as AdamW optimizer.

We decided to continue using mBERT in $\text{UScore}_{\text{WRD}}$ for two reasons. Firstly, we want $\text{UScore}_{\text{WRD}}$ to remain directly comparable to XMoverScore which is based on mBERT. Secondly, in our own experience, vanilla mBERT is very robust in terms of layer choice, especially compared to vanilla XLM-R. Intuitive choices like the first or last layer work very well for a lot of problems. This is an important property for unsupervised metrics, since we can't easily justify a (supervised) hyperparameter search in an unsupervised setting to determine the best layer or even a linear combination of those.

E Supplementary Data and Results

Table 3 shows examples of pseudo-parallel data obtained with $\text{UScore}_{\text{WRD}}$ and $\text{UScore}_{\text{SNT}}$. Tables 4, 5, 6, and 7 show segment-level Pearson's r correlations with human judgments on WMT-16, WMT-17, MLQE-PE, as well as WMT-MQM and Eval4NLP, respectively. Table 8 provides additional statistics for each dataset.

A surprising finding is the poor performance of $\text{UScore}_{\text{SNT}}$ and DistilScore on the German-English language pair in Table 7. It is well known that high-resource language directions, such as English-German, can be affected by a lack of low-quality translations (Fomicheva et al., 2022), and with only high-quality translations available, there is little variation in the scores, which makes a meaningful assessment of correlation difficult (Specia et al., 2020). Further, since both $\text{UScore}_{\text{SNT}}$ and DistilScore are based on sentence embeddings of XLM-R, and sentence embedding-based metrics are known to be a bit worse on average than their word embedding-based counterparts, we believe that both aspects combined could be the root cause for this.

Case	Source	Target
Top-WRD	Uruguay belegt mit vier Punkten nur Platz Sieben.	Russia was second with four gold and 13 medals.
Top-WRD	Soweit lautet zumindest die Theorie.	That, at least, is the theory.
Rnd-WRD	Die USA stellen etwa 17.000 der insgesamt 47.000 ausländischen Soldaten in Afghanistan.	Currently, there are about 170,000 U.S. troops in Iraq and 26,000 in Afghanistan.
Rnd-WRD	"Das ist eine schwierige Situation", sagte Kaczynski.	"It seemed like a ridiculous situation," Vanderjagt said.
Top-SNT	Die Wahlen für ein neues Parlament sollen dann Anfang Januar stattfinden.	Parliamentary elections are to be held by January.
Top-SNT	Anzeichen für die Blauzungenkrankheit sind Fieber, Entzündungen und Blutungen an der Zunge der Tiere.	Contact with the creatures can cause itching, rashes, conjunctivitis and, in some cases, breathing problems.
Rnd-SNT	Riesen-Wirbel an der Universität Zagreb: An der wirtschaftlichen Fakultät und am Institut für Verkehrsstudien durchsuchen Polizisten die Büros von Dozenten.	Those attending the Soil Forensics International Conference work in the fields of science, policing, forensic services as well as private industries.
Rnd-SNT	Frankfurt soll WM-Finale der Frauen ausrichten	The women's tournament gets underway on Sunday.

Table 3: Pseudo-parallel data obtained via $\text{USCORE}_{\text{WRD}}$ and $\text{USCORE}_{\text{SNT}}$; top and random sentence pairs. The mined sentences are semantically similar, but contain factuality errors (e.g., have wrong places or numbers in hypotheses).

	Metric	de-en	en-ru	ru-en	ro-en	cs-en	fi-en	tr-en
TYPE-1/2 Supervised	MONOTRANSQUEST	61.65	66.69	63.32	62.36	67.67	68.33	62.71
	COMET-QE	65.73	71.91	69.71	66.40	67.98	64.83	53.73
	BLEU	45.39	55.08	46.33	47.09	53.80	39.92	47.15
TYPE-3 Supervised	XMOVERSCORE (UMD)	43.46	62.16	**60.52**	47.88	58.83	43.52	33.34
	XMOVERSCORE (CLP)	45.29	63.58	56.12	54.24	58.89	51.40	42.14
	SENTSIM (BERTSCORE)	48.49	50.37	57.89	55.06	59.08	46.66	**45.47**
	SENTSIM (WMD)	47.78	48.49	56.19	54.48	56.87	46.00	44.84
	DISTILSCORE	40.62	41.21	43.16	51.22	49.51	39.65	41.14
Unsupervised	$\text{USCORE}_{\text{WRD}}$	48.94	60.97	59.09	56.06	57.15	53.76	41.15
	$\text{USCORE}_{\text{SNT}}$	30.48	39.73	35.31	44.18	40.80	31.30	30.62
	$\text{USCORE}_{\text{WRD} \oplus \text{SNT}}$	**50.37**	62.92	60.49	60.07	**59.25**	52.93	44.41
	$\text{USCORE}^{+}_{\text{WRD}}$	48.97	60.57	58.35	56.23	56.44	53.78	41.30
	$\text{USCORE}^{+}_{\text{WRD} \oplus \text{SNT}}$	50.29	62.37	59.43	**60.71**	58.47	51.48	44.19
	$\text{USCORE}^{++}_{\text{WRD}}$	44.50	61.45	53.50	56.03	52.95	54.08	40.65
	$\text{USCORE}^{++}_{\text{WRD} \oplus \text{SNT}}$	47.27	**64.14**	56.40	60.36	56.29	**55.62**	44.85

Table 4: Segment-level Pearson's *r* correlations with human judgments on the WMT-16 dataset.

	Metric	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en
TYPE-1/2 Supervised	MONOTRANSQUEST	60.93	63.54	65.33	72.01	59.56	75.91	68.03
	COMET-QE	68.99	69.34	72.83	64.75	69.06	68.43	68.01
	BLEU	41.22	41.29	56.48	39.28	45.99	53.06	51.75
TYPE-3 Supervised	XMOVERSCORE (UMD)	41.72	51.19	56.61	56.11	47.24	46.52	57.73
	XMOVERSCORE (CLP)	47.76	50.04	62.22	63.95	48.79	52.97	59.88
	SENTSIM (BERTSCORE)	49.90	52.26	57.85	57.42	**55.10**	56.84	59.59
	SENTSIM (WMD)	47.62	50.42	56.59	56.91	53.42	56.24	58.86
	DISTILSCORE	46.42	45.64	54.03	55.51	54.13	54.04	50.89
Unsupervised	$\text{USCORE}_{\text{WRD}}$	46.70	52.71	61.91	59.22	49.10	50.06	54.95
	$\text{USCORE}_{\text{SNT}}$	38.89	41.92	39.77	48.95	37.27	48.83	43.10
	$\text{USCORE}_{\text{WRD} \oplus \text{SNT}}$	**50.67**	**56.50**	61.86	64.54	53.17	57.31	58.80
	$\text{USCORE}^{+}_{\text{WRD}}$	44.55	52.96	60.96	59.02	48.82	49.70	54.75
	$\text{USCORE}^{+}_{\text{WRD} \oplus \text{SNT}}$	48.62	56.31	60.44	63.62	54.19	**57.54**	**60.28**
	$\text{USCORE}^{++}_{\text{WRD}}$	45.41	48.00	61.33	60.64	49.65	47.06	50.00
	$\text{USCORE}^{++}_{\text{WRD} \oplus \text{SNT}}$	49.27	52.20	**63.10**	**66.64**	53.70	54.93	55.62

Table 5: Segment-level Pearson's *r* correlations with human judgments on the WMT-17 dataset.

	Metric	en-de	en-zh	ru-en	ro-en	et-en	ne-en	si-en
TYPE-1 Supervised	MONOTRANSQUEST	41.85	45.76	76.76	88.81	73.19	75.70	61.90
	COMET-QE	36.03	30.70	49.33	64.95	63.45	57.46	47.85
TYPE-3 Supervised	XMOVERSCORE (UMD)	16.56	16.48	28.07	65.83	53.95	38.23	18.81
	XMOVERSCORE (CLP)	25.59	20.25	20.31	57.34	**58.46**	25.15	16.74
	SENTSIM (BERTSCORE)	6.15	22.23	47.30	**78.55**	55.09	**57.09**	**51.14**
	SENTSIM (WMD)	3.86	22.62	**47.46**	77.72	54.60	57.00	49.79
	DISTILSCORE	12.96	**28.68**	45.34	76.57	51.16	46.73	41.76
Unsupervised	$\text{USCORE}_{\text{WRD}}$	24.53	21.58	16.66	60.30	54.83	28.62	11.38
	$\text{USCORE}_{\text{SNT}}$	13.62	13.27	39.09	66.50	41.75	46.83	41.49
	$\text{USCORE}_{\text{WRD} \oplus \text{SNT}}$	24.67	21.63	27.83	71.48	57.62	45.70	35.25
	$\text{USCORE}^{+}_{\text{WRD}}$	23.78	21.40	15.18	60.39	56.27	27.34	12.22
	$\text{USCORE}^{+}_{\text{WRD} \oplus \text{SNT}}$	22.00	21.41	27.71	73.65	57.94	47.27	39.91
	$\text{USCORE}^{++}_{\text{WRD}}$	26.17	27.16	0.75	32.92	52.24	12.79	14.30
	$\text{USCORE}^{++}_{\text{WRD} \oplus \text{SNT}}$	**26.37**	27.26	9.29	57.70	56.81	31.24	30.38

Table 6: Segment-level Pearson's *r* correlations with human judgments on the MLQE-PE dataset.

		WMT-MQM		Eval4NLP	
	Metric	en-de	zh-en	de-zh	ru-de
TYPE-1/2 Supervised	MONOTRANSQUEST	31.38	53.07	34.66	49.76
	COMET-QE	40.94	50.84	33.51	46.49
	BLEU	17.94	21.58	—	—
TYPE-3 Supervised	XMOVERSCORE (UMD)	12.72	27.41	10.85	48.35
	XMOVERSCORE (CLP)	12.89	28.29	21.61	52.25
	SENTSIM (BERTSCORE)	1.03	26.39	-7.71	**60.91**
	SENTSIM (WMD)	-0.23	24.70	-11.54	60.42
	DISTILSCORE	-2.66	7.06	6.57	51.22
Unsupervised	$\text{USCORE}_{\text{WRD}}$	18.13	**28.43**	29.29	44.63
	$\text{USCORE}_{\text{SNT}}$	-5.93	4.99	0.86	39.91
	$\text{USCORE}_{\text{WRD} \oplus \text{SNT}}$	13.94	25.17	24.66	55.08
	$\text{USCORE}^{+}_{\text{WRD}}$	19.83	26.46	29.99	42.59
	$\text{USCORE}^{+}_{\text{WRD} \oplus \text{SNT}}$	13.72	21.7	22.64	53.8
	$\text{USCORE}^{++}_{\text{WRD}}$	**20.19**	17.80	**36.51**	14.32
	$\text{USCORE}^{++}_{\text{WRD} \oplus \text{SNT}}$	17.18	16.97	34.00	30.09

Table 7: Segment-level Pearson's *r* correlations with human judgments on the WMT-MQM and Eval4NLP datasets.

Dataset	Pairs	Tokens	Type
WMT-16	560	9886	CA
WMT-17	560	11189	CA
MLQE-PE	1000	15263	CLDA
Eval4NLP	1295	25975	CLDA
WMT-MQM	17090	549835	MQM

Table 8: Statistics of our used datasets averaged over language pairs. For each dataset we report the number of sentences we evaluated on, the amount of tokens in these, and the type of human annotation. The number of tokens refers to source sentences.