Crosslingual Semantic Evaluation of Machine Translation

Anonymous NAACL-HLT 2021 submission

Abstract

Machine translation (MT) is currently evaluated in one of two ways: in a monolingual fashion, by comparing the system output to one or more human reference translations, or in a trained crosslingual fashion, by building a supervised model to predict quality scores from human-labeled data. In this paper we propose a more cost-effective, yet well performing alternative **SentSim**: relying on strong pretrained multilingual word and sentence representations, we directly compare the source to the machine translated sentence, thus avoiding the need for both reference translations and training data. The metric builds on state-of-the-art embedding-based approaches – namely BERTScore and Word Mover's Distance – by incorporating a notion of sentence semantic similarity. By doing so, it achieves better correlation with human scores on different MT datasets. We also show that (i) it outperforms these metrics in the standard monolingual setting (MT-reference translation), (ii) the source-MT comparison follows the same trend as that of MT-reference variant, (iii) it performs on pair with glass-box approaches quality estimation, which use MT model information.

1 Introduction

Automatically evaluating machine translation (MT) as well as other language generation tasks has been investigated for decades, with substantial progress in recent years due to the advances of pretrained contextual word embeddings. The general goal of such evaluation metrics is to to estimate the semantic equivalence between the input text (e.g. a source sentence or a document) and an output text that has been modified in some way (e.g. a translation or summary), as well as the general quality of the output (e.g. fluency). As such, by definition metrics should perform some form of input-output comparison.

However, this direct comparison has proved hard in the past because of the natural differences between the two versions (such as different languages). Instead, evaluation metrics have resorted to comparisons against one or more correct outputs produced by humans, *a.k.a.* reference texts, where comparisons at the string level are possible and straightforward. A multitude of evaluation metrics have been proposed following this approach, especially for MT, the application we focus on in this paper. These include the famous BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2005) for machine translation, ROUGE (Lin, 2004) for summarization, and CIDER (Vedantam et al.) for image captioning. These traditional metrics are based on simple word or n-gram matching mechanisms or slight relaxations of these (e.g. synonyms) are computationally efficient, but suffer from various limitations.

In order to overcome the drawbacks of the traditional string-based evaluation metrics, recent work (Williams et al., 2018; Bowman et al., 2015; Echizen'ya et al., 2019; Cer et al., 2017; Echizen'ya et al., 2019) has investigated metrics that perform comparisons in the semantic space rather than at the surface level. Notably, applications of Word Mover's Distance (WMD; Kusner et al., 2015), such as WMDo (Chow et al., 2019), VIFIDEL (Madhyastha et al., 2019) or moverscore (Zhao et al., 2019) , which both compute similarity based on continuous word embeddings using pretrained representations. These have been shown to consistently outperform previous metrics on various language generation evaluation tasks.

However, these metrics have two limitations: (i) they still rely on reference outputs, which as previously discussed come with their own issues; (ii) these are bag-of-embeddings approaches which capture semantic similarity at the word level, but are unable to capture the meaning of the sentence or text as a whole, including correct word order.

In this paper, focusing on MT, to address these limitations we first posit that evaluation can be done directly from comparing the source to the machine translation using multilingual pretrained embeddings, such as multilingual BERT, avoiding the need of reference translations. We note that this is different from quality estimation (QE) metrics (Specia et al., 2013; Shah et al., 2015) , which also compare source and machine translated texts directly, but assume an additional step of supervised learning against human labels for quality. Second, we introduce Sentence Semantic Similarity (SSS) , an additional component to be combined with bag-of-embeddings distance metrics such as BERTScore. More specifically, we propose to explore semantic similarity at the sentence level – based on sentence embeddings (Reimers and Gurevych, 2019, 2020; Thakur et al., 2020) – and linearly combine it with the existing metrics that uses the contextual word embeddings. By doing so, the resulting metrics have access to a higher-level semantics of the two versions of the text, with improved performance. The combination is a simple weighted sum, which require no training data.

As a motivational example, consider the case in Table 1, from the WMT-17 Metrics task (Zhang et al., 2020a). When faced with MT sentences that contain a negated version of the reference (MTs 3 and 4), token-level metrics such as BERTScore and WMD cannot correctly penalize these sentences since they match representations of words in both versions without a full understanding of the semantics of the sentences. As a consequence, they return a high score for these incorrect translations, higher than the score for correct paraphrases of the reference (MTs 1 and 2). Sentence similarity, on the other hand, correctly captures this mismatch in meaning, returning relatively lower scores for Translations 3 and 4, but may be considered to harsh as the remaining of the sentence has the same meaning. The combination of these two metrics (last column) balances between these two sources of information and, as we will later show in this paper, has higher correlation with human scores.

Our **main contributions** are:

1. We investigate and show the effectiveness of linearly combining sentence-level semantic similarity with different metrics using word-level semantic similarity. The resulting combined metrics, SentSim, consistently achieve higher Pearson correlation with human judgements of translation quality than both word and sentence similarity alone.

2. We show, for the first time, that these metrics can be effective for comparing system-generated sentences not only with reference translations, but also with source sentences, in a crosslingual fashion.

3. Our SentSim metrics outperform existing metrics' performances on various MT datasets in the monolingual and crosslingual cases.

2 Related Work

Various natural language generation tasks, including machine translation, image captioning, among others, generally produce sentences as output. These are evaluated either manually or automatically by comparison with one or multiple reference sentences. A multitude of metrics have been proposed for the latter, which perform comparisons at various granularity levels, from characters to embedding vectors. The goal of such metrics is to replace human judgements. In order to understand how well they fair at this task, metrics are evaluated by how similar their scores is to human assigned judgements on held-out datasets. For absolute quality judgements, Pearson correlation is the most popularly used metric used metric for such a comparison (Mathur et al., 2020).

Recent studies have shown that the new generation of automatic evaluation metrics that instead of lexical overlap are based on word semantics using continuous word embedding, such as BERT (Devlin et al., 2019), ElMo (Peters et al., 2019), XLNet (Yang et al., 2019) or XLM-Roberta (Conneau et al., 2019), have significantly higher Pearson correlation with the human judgements when comparing reference sentences with system generated sentences. Zhang et al. (2019) introduced **BERTscore**, an automatic evaluation metric based on contextual word embeddings, and tested it for text generation tasks such as machine translation and imaging captioning, using embeddings including BERT, XLM-Roberta, and XLNet (more details in Section 3.2). Mathur et al. (2019) present supervised and unsupervised metrics which are based on BERT embeddings for improving machine translation evaluation. Zhao et al. (2019) introduces **moverscore**, a metric which generates high-quality evaluation results on a number of text generation tasks including summarization, machine translation, image captioning, and data-to-text generation, using BERT embeddings. Clark et al. (2019) present semantic metrics for text summarization based on the **sentence**

		BERTScore	SSS	SSS + BERTScore
REF	We have made a complete turnaround.			
MT1	We did a complete turnaround.	0.7975	0.9578	0.8111
MT2	We made a turnaround.	0.7748	0.8898	0.7427
MT3	We have not made a complete turnaround.	0.8296	0.3878	0.4832
MT4	We have made an incomplete turnaround.	0.8318	0.4431	0.5107

Table 1: An example from the WMT-17 dataset. Given the Reference sentence, BERTScore assigns higher similarity to its negated versions (MTs 3 and 4) than to semantically similar variants (MTs 1 and 2). Contrarily, semantic sentence similarity gives a very low score to MTs 3 and 4. Their combination provides a more balanced score.

mover's similarity and ELMo embeddings. Chow et al. (2019) introduce fluency-based word mover's distance (**WMDo**) metric for machine translation evaluation using word2vec embeddings (Mikolov et al., 2013). Lo (2019) presents **Yisi**, a unified automatic semantic machine translation quality evaluation and estimation metric, using BERT embeddings.

There is also a bulk of work on metrics that take a step further to optimize their scores using machine learning algorithms trained on human scores for quality (Sellam et al., 2020; Ma et al., 2017). They often perform even better, but the reliance on human scores for training, in addition to reference translations at inference time, makes them less applicable in practice. A separate strand of work that relies on contextual embeddings is that of Quality Estimation (Moura et al., 2020; Fomicheva et al., 2020a; Ranasinghe et al., 2020; Specia et al., 2020). These are also trained on human judgements of quality, but machine translations are compared directly to the source sentence rather than against reference translations.

In addition to embeddings for words, embeddings for full sentences have been shown to work very well to measure semantic similarity. These are transformer models that are specifically trained for capturing sentence semantic meanings using BERT, Roberta, and XLM-Roberta embeddings (Reimers and Gurevych, 2019; Reimers and Gurevych, 2020; Thakur et al., 2020) and provides state-of-art performance pretrained models for more than 100 languages.[1]

In this paper we take inspiration from these lines of previous work to propose untrained metrics that combine word and sentence semantic similarity and show that this can be effective for both MT-reference and source-MT comparisons.

[1] https://github.com/UKPLab/sentence-transformers

3 Method

In this section, we will first describe in more detail the metrics that we have used in our experiments, namely semantic sentence cosine similarity, WMD and BERTScore. Second, we will present our simple linear combination method between these.

3.1 Word Mover's Distance (WMD)

Kusner et al. (2015) present word mover's distance (WMD) metric, a special case of Earth mover's distance (Rubner et al., 2000), computing the semantic distance between two text documents by aligning semantically similar words and capturing the word traveling flow between the similar words utilizing the vectorial relationship between their word embeddings (Mikolov et al., 2013). WMD has been proven to generate consistently high-quality results for the tasks of finding text similarity and text classification (Kusner et al., 2015). A text document is represented as a vector D, where each element is denoted as the normalized frequency of a word in the document such that:

$$D = [d_1, d_2, ..., d_n]^T \tag{1}$$

where $d_i = c_i / \sum_j^n c_j$ and c_i is the i^{th} word appears c_i times in a given text document. Assuming there are two given words from different text document denoted as x_i and x_j, then the euclidean distance in the embedding for the two words is defined as:

$$c(i, j) = \|x_i - x_j\|_2 \tag{2}$$

where $c(i, j)$ is defined as the "word traveling cost" from x_i in one document to x_j in the other document. Now, assuming there are two documents, one is the source document denoted as A where the word x_i belongs to, and another one is the target document denoted as B where the word x_j belongs to. A flow matrix T is defined in which every element is denoted as T_{ij}, suggesting the number of times the word x_i in document A moves to the word x_j in document B. Then, the value of the flow matrix is normalized based on the total count of words in the vocabulary such that:

$$\sum_j T_{ij} = d_i, \sum_i T_{ij} = d_j \quad (3)$$

The semantic distance calculated by WMD can be then defined as follows:

$$WMD = \min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij} c(i,j) \quad (4)$$

WMD, which represents the semantic distance between two text documents, can thus be computed by optimizing values in the flow matrix T. In other words, WMD corresponds to the minimal semantic distance to move semantically similar words from one text document to another.

3.2 BERTScore

BERTScore (Zhang et al., 2020b) was designed to evaluate semantic similarity between sentences in the same language, namely a reference sentence and a machine-generated sentence. Assume a reference sentence is denoted as $x = (x_1, ..., x_k)$ and a candidate sentence denoted as $\hat{x} = (\hat{x_1}, ..., \hat{x_k})$, BERTScore uses contextual embeddings such as BERT (Devlin et al., 2019) or ELMo (Peters et al., 2019) to represent word tokens in the sentences. It computes the word matching between the reference and candidate sentence using cosine similarity, which can be optionally reweighted by the inverse document frequency scores (IDF) of each word. BERTScore matches each word token x in reference sentence to the closest word token $\hat{x}$ in candidate sentence for computing recall, and matches each word token $\hat{x}$ in candidate sentence to the closets word token x in reference sentence for computing precision. It combines recall with precision to produce the F1 score. However, only recall is used for evaluation in most cases, which is defined as follows:

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x_j} \in \hat{x}} x_i^T \hat{x_j} \quad (5)$$

In essence, BERTScore can be viewed as a hard word alignment given a pair of sentences using contextual embeddings, in which each word is aligned to one other word, the closest in the embedding space according to the cosine distance between their vectors.

3.3 Semantic Sentence Similarity (SSS)

A commonly used method to measure sentence similarity is using he cosine distance between the two vectors summarising the sentences:

$$cos(\theta) = (A \cdot B)/(\|A\|\|B\|) \quad (6)$$

where θ is the cosine of the angle between two vectors, and A and B are the vectors representing the two sentences. The higher the value obtained through cosine similarity between two sentences vectors based on the pretrained sentence representation (Reimers and Gurevych, 2019; Reimers and Gurevych, 2020; Thakur et al., 2020), the stronger their similarity will be.

In order to bring the notion of semantic similarity to token similarity metrics, we linearly combine the sentence cosine similarity using semantically fine-tuned sentence embedding with the metrics using contextual word embeddings. Assume that the value generated from semantic sentence cosine similarity metric is denoted as J and the value generated from another metric which is linearly combined with sentence cosine similarity is denoted as L. Our linear combination method, namely SentSim, follows:

$$SentSim(J, L) = exp(J) + exp(L) \quad (7)$$

where J and L are scaled to the range between 0 and 1.

If metric L is negative correlated to our target, we give it $exp(1 - L)$, such as WMD metric. Otherwise, it uses $exp(L)$ like the other positively correlated metric. We use exp during the linear combination for mapping the scaled combined score (0 to 1) to a reasonable range (1 to e) following the similar example in the work of (Clark et al., 2019).

We explore two kinds of linear metric combinations, including **(SSS + WMD)**, **(SSS + BERTScore)**. The performance of these metric combinations is presented in Section 5. We note that we also investigated the linear combination between Sentence Mover's Distance (Zhao et al., 2019) and token-level metrics. However, we found that the results ae poorer than with SSS, as shown in the Appendix A.2.

4 Experiment Setup

In this section, we describe the two types of experimental scenarios, monolingual evaluation and

	SRC - MT			REF - MT		
Metrics	en-de	en-fr	Avg	de-de	fr-fr	Avg
WMD	0.360	0.319	0.340	0.492	0.425	0.459
SSS	0.483	0.449	0.466	0.487	0.446	0.467
BERTScore	0.335	0.291	0.313	0.434	0.352	0.393
SSS + WMD	**0.508**	**0.477**	**0.492**	**0.546**	**0.501**	**0.524**
SSS + BERTScore	0.486	0.434	0.460	0.527	0.462	0.494

Table 2: Pearson correlation with human score for Multi-30K with XLM-Roberta-Base. In the the MT-REF case, we uses XLM-Roberta-Base embeddings to evaluate German to German and French to French monolingual tasks.

	SRC-MT									
Metrics	de-en	zh-en	fi-en	lv-en	ru-en	cs-en	en-ru	en-zh	tr-en	Avg
WMD	0.366	0.501	0.373	0.373	0.308	0.267	0.404	0.408	0.350	0.372
SSS	0.456	0.514	0.540	0.555	0.541	0.464	0.505	0.458	0.540	0.508
BERTScore	0.409	0.510	0.414	0.402	0.337	0.319	0.434	0.446	0.382	0.406
SSS + WMD	0.504	0.594	0.566	0.569	0.534	0.476	0.538	0.513	0.562	0.540
SSS + BERTScore	**0.523**	**0.600**	**0.578**	**0.574**	**0.551**	**0.499**	**0.553**	**0.531**	**0.569**	**0.553**

Table 3: Pearson correlation with human score of WMT-17 with XLM-Roberta-Base in the case of SRC-MT setting.

crosslingual evaluation, as well as the three datasets and pretrained embeddings we used.

4.1 Task Scenarios

The first evaluation setting we experimented with is the standard monolingual evaluation task scenario (MT-REF), which takes reference sentences and machine generated sentences in the same language as input. The second one is the crosslingual evaluation task scenario (SRC-MT), which directly assesses the similarity between source sentences and machine generated sentences in different languages. We then compute our combined metrics for each task scenario separately.

4.2 Datasets

Multi-30K[2] is a multilingual (English-German (en-de) and English-French (en-fr)) image description dataset. We use the 2018 test set, in which each language contains more than 2K sentence tuples, including source sentences, reference sentences, machine generated sentences, and the corresponding human judgement scores in an (0-100) continuous range. Therefore, this dataset can be used for both crosslingual and monolingual task scenarios.

WMT-17[3] segment-level is a dataset containing 14 different language pairs from news task. We used all seven to-English task German-English (de-en), Chinese-English (zh-en), Latvian-English (lv-en), Czech-English (cs-en), Finnish-English (fi-en), Russian-English (ru-en), Turkish-English (tr-en) and two from-English task English-Russian (en-ru), English-Chinese (en-zh). As each language has 560 sentence tuples, where each tuple has one source sentence, one reference sentence and multiple system generated sentences, in addition to a human score varying from 0 to 100. WMT-17 can be used in both monolingual evaluation and crosslingual task scenarios, and is our main experimental data. This was the last year in which the WMT Metrics evaluation campaign used absolute human scores for metric comparison. In more recent datasets, pairwise judgements are used instead. While such pairwise judgements are useful to assess metrics ability to rank different MT systems, they are not applicable to assess metrics in their ability to estimate quality in absolute terms, which is what we are interested in.

WMT-20[4] comes from the quality estimation task, where participants are expected to directly predict the translation quality between source sentences and machine generated sentences without using reference sentences. This dataset has seven language pairs: Sinhala-English (si-en), Nepalese-English (ne-en), Estonian-English (et-en), English-German (en-de), English-Chinese (en-zh), Romanian-English (ro-en), Russian-English (ru-en). Each language pair contains 1K tuples with source and machine generated sentences, as well as human judgements in 0-100. Therefore, with this dataset we can only perform crosslingual evaluation.

[2] https://github.com/multi30k/dataset
[3] http://www.statmt.org/wmt17/
[4] http://www.statmt.org/wmt20/quality-estimation-task.html

4.3 Embeddings

In our experiments, XLM-Roberta-Base and Roberta-Large are used for crosslingual and monolingual assessments, respectively. For each language model, we also consider embeddings at the word level and sentence level individually. The reason why we use XLM-Roberta-Base is because XLM-Roberta-Base outperforms multi-lingual Bert (mBERT) Base significantly, as the multilingual masked language model, mentioned by (Conneau et al., 2019). As described in the (Reimers and Gurevych, 2020), XLM-Roberta performs better than mBERT and Distilled mBERT in crosslingual semantic textual similarity (STS) sentence embedding tasks (Cer et al., 2017) as well. For crosslingual word embeddings, we replaced the traditional Word2Vec previously used (e.g. Chow et al., 2019) with XLM-Roberta-Base embedding and provided the same embedding for BERTScore as well. However, for the semantic sentence embedding, we used XLM-Roberta-Base embedding from Sentence Transformer, which were trained on SNLI (Bowman et al., 2015) + MultiNLI (Williams et al., 2018) and then fine-tuned on the STS benchmark training data. These sentence embeddings have been shown to provide good representations of the semantic relationship between two documents, especially in the crosslingual scenario, even though they had not yet been tested for machine translation evaluation. Without using semantic embeddings, the performance of SSS is not consistent across different languages pairs given the experiment datasets (see Appendix A.2). The monolingual word embeddings and semantic sentence embeddings are similar to the ones used in crosslingual tasks but are based on the Roberta-Large model, as BERTScore (Zhang et al., 2020a) reaches the state-of-art performance in the WMT-17 monolingual task by using Roberta-Large embeddings.

5 Results

The evaluation results are presented in this section. Our project code and experimental data will be publicly released.

5.1 SRC-MT Setting

From the Table 2, we can observe the Pearson correlation of our metrics comparing the source sentences to machine generated sentences, both single metrics and their combinations in the Multi-30K dataset. The result reveals that **SSS + WMD** outperforms individual metric and the other compound metrics. It is clear that SSS is better than both WMD and BERTScore, with WMD outperforming BERTScore in this specific crosslingual task.

In the Table 3, the benefit of SSS becomes even more evident. It again outperforms WMD and BERTScore, with BERTScore also significantly outperforming WMD in this case, and **Semantic Sentence Simolarity + BERTScore** showed the best and more stable performance for all language pairs in WMT-17 dataset. This can be also seen in Figure 1. By observing the combined met-

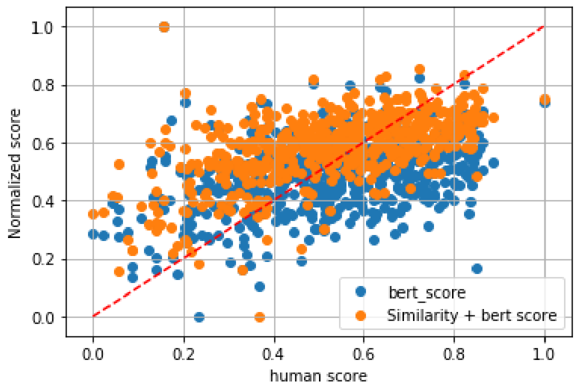

Figure 1: BERTScore against Sentence Sentence SSS + BERTScore for lv-en in WMT-17 SRC-MT case.

ric performance difference in dataset Multi-30K and WMT-17, we believe this fact happens because sentence length differs in datasets Multi-30k and WMT-17, and the sentence length in Multi-30K, around 12-14 words, normally is shorter than which in WMT-17. Because WMD optimizes the word alignment globally for the whole sentence instead of optimizing word alignment locally like BERTScore, the performance of WMD would be better than BERTScore when the evaluating sentence length is short, but will be penalized when the sentence length gets long. Therefore, the performance of **SSS + WMD** is better than that of **SSS + BERTScore** in dataset Multi-30K but lower than that of **SSS + BERTScore** in dataset WMT-17. SSS also outperforms to WMD and BERTScore in WMT-20 dataset, as Table 4 shows. **SSS + BERTScore** reaches the state-of-art in three out of seven language pairs and is the best metric in comparison with BERTScore or WMD alone. The metrics that outperform **SSS + BERTScore** for three language pairs require multiple passes of the neural machine translation decoder to score or generate multiple translations (D-TP and D-Lex-Sim, re-

Metrics	SRC-MT							
	ne-en	en-de	et-en	en-zh	ro-en	si-en	ru-en	Avg
Leaderboard baseline	0.386	0.146	0.477	0.190	0.685	0.374	**0.548**	0.322
D-TP	0.558	0.259	0.642	0.321	0.693	0.460	—	0.489
D-Lex-Sim	**0.600**	0.172	**0.612**	0.313	0.663	**0.513**	—	0.479
WMD	0.361	0.456	0.463	0.251	0.647	0.308	0.315	0.400
SSS	0.313	0.330	0.481	0.401	0.694	0.404	0.441	0.438
BERTScore	0.357	0.459	0.460	0.260	0.673	0.309	0.320	0.405
SSS + WMD	0.390	0.472	0.553	**0.427**	0.724	0.426	0.476	0.495
SSS + BERTScore	0.392	**0.484**	0.553	**0.427**	**0.727**	0.426	0.475	**0.498**

Table 4: Pearson correlation with human score for the WMT-20 dataset with XLM-Roberta-Base in the case of SRC-MT. Metrics like D-TP and D-Lex-Sim (Fomicheva et al., 2020b) are unsupervised metrics which show good performance in the WMT-20 quality estimation shared task, while Leaderboard baseline is a supervised model provided by the organizers that uses training data to finetune pretrained representations.

Metrics	MT-REF							
	de-en	zh-en	fi-en	lv-en	ru-en	cs-en	tr-en	Avg
MEANT 2.0	0.565	0.639	0.687	0.586	0.607	0.578	0.596	0.608
WMD_o (Word2Vec)	0.531	0.595	0.689	0.505	0.562	0.513	0.561	0.565
WMD_o (BERT)	0.546	0.623	0.710	0.543	0.585	0.531	0.637	0.596
WMD	0.730	0.769	0.827	0.736	0.733	0.698	0.770	0.752
SSS	0.612	0.653	0.730	0.703	0.700	0.622	0.654	0.668
BERTScore	0.745	0.775	0.833	0.756	0.746	0.710	0.751	0.759
SSS + WMD	0.755	0.779	0.847	0.781	0.786	0.731	0.781	0.780
SSS + BERTScore	**0.770**	**0.785**	**0.860**	**0.792**	**0.796**	**0.746**	**0.782**	**0.790**

Table 5: Pearson correlation with human score for the WMT-17 dataset with Roberta-Large in the case of MT-REF (to English). MEANT 2.0 (Lo, 2017) was the winning metric that year, while WMD_o is from (Chow et al., 2019) and uses word2vec embeddings instead.

spectively), or require supervised machine learning (Leaderboard baseline [5]).

5.2 MT-REF Setting

In the machine generated sentence to reference sentence case, as Table 2 shows, **SSS + WMD** achieves the best result in the monolingal Multi-30K tasks for both German to German and French to French using XLM-Roberta-Base embedding. In the standard setting where we compare sentences in a monolingual fashion, as we can observe from Table 5, **SSS + BERTScore** is the best metric, although if taken independently, the performance of SSS is not as good here as that of WMD or BERTScore. The two variants of the combined metric still outperform any metric on their own, and reach the state-of-art results in this dataset. It can also observed from Table 5 that WMD_o with word2vec is far behind than that with BERT embedding or our WMD with Roberta-Large. It indicates that the importance of using the pretrained contextual embedding as the representation of tokens. A visual example of correlation plots can be seen in Figure 2.

Generally, the metrics' performances in the case of SRC-MT are much lower than that of MT-REF. This is to be expected and can be attributed to a few reasons behind this. First, the models' embeddings are not the same in these two cases. In the case of MT-REF, monolingual embeddings are used, which are known to be stronger; however these cannot be used in the case of SRC-MT evaluation, where crosslingual embeddings are used. In fact, the crosslingual embeddings have been trained on more than 100 languages. Therefore, the performance of crosslingual embedding is lower than that of monolingual embedding, which is designed specific for one language. Second, the size of pretrained model for the case of MT-REF (Roberta-Large) is much larger than that of SRC-MT (XLM-Roberta-Base). Due to lack of semantic sentence embedding from XLM-Roberta-Large currently, we provided the performance of MT-REF (Roberta-Base) about WMT-17 in appendix A.4. Finally, there is no specific alignment or mapping for the crosslingual tokens or sentences in the embedding space, a fact which can be another reason causing the metric performance dropping in the case of SRC-MT.

[5] https://competitions.codalab.org/competitions/24447

			BERTScore	SSS	SentSim
E1	REF	The food tastes good.			
E1	MT1	The food tastes not good.	0.954	0.778	0.821
E1	MT2	The food tastes not bad.	0.948	0.948	0.950
E2	REF	President Barack Obama also backs the proposal.			
E2	MT1	President Obama also supported this proposal.	0.8419	0.954	0.844
E2	MT2	Supported President Obama also this proposal.	0.4604	0.625	0.405
E3	REF	She is recovering, and police are still searching for a suspect.			
E3	MT1	She is recovering, and police are searching for a suspect.	0.984	0.903	0.912
E3	MT2	Police searched for a suspect, and she recovered.	0.911	0.688	0.713
E4	SRC	The food tastes good.			
E4	MT1	这食物味道好.	0.882	1.000	0.958
E4	MT2	好道味物食这. (word order shuffled)	0.207	0.682	0.309

Table 6: Examples from various datasets comparing BERTScore, SSS and SentSim(SSS + BERTScore).

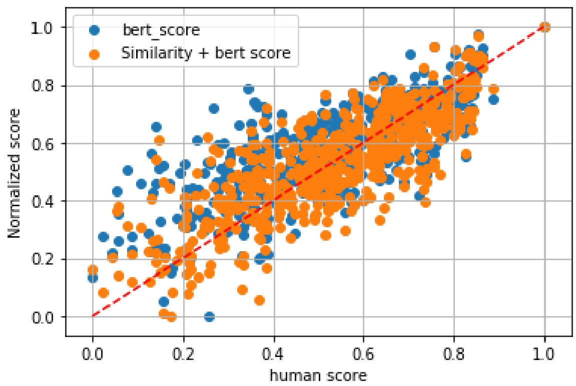

Figure 2: BERTScore against SSS + BERTscore for lv-en in WMT-17 MT-REF case.

5.3 Discussion

For illustration purposes, Table 6 shows a few cases where SSS performs better than token-level metric because it adds the notion of sentence meaning and where, as a consequence, SentSim performs better (examples E1 and E2). It also show cases where SSS is too sensitive to semantic changes (example E3). SSS also performs well in the SRC-MT case (example E4). Here, interestingly the second machine translation has very different and incorrect word order, and the token-level metric (BERTScore) has very low performance compared to SSS, but both token-level and SSS metrics capture the incorrect word order. The combined metric (SentSim), therefore, is very robust.

6 Conclusions

In this paper, we propose to combine sentence-level and token-level evaluation metrics in an unsupervised way, where no human-labelled training data is required. In our experiments on a number of standard datasets, we demonstrate that this combination is more effective for MT evaluation than the state-of-the-art token-level metrics, substantially outperforming these as well as sentence-level semantic metrics on their own. The sentence level metric captures the overall, high-level semantic similarity, which complements the token level semantic information.

We also show that this combination approach can be applied both in the standard monolingual evaluation setting, where machine translations are compared to reference translations, and in the a crosslingual evaluation setting, where reference translations are not available and machine translations are directly compared to the source sentences.

In future work, we will aim to improve the crosslingual metric and explore other types of multilingual embeddings for better mapping across different languages.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. WMDO: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day*

1), pages 494–500, Florence, Italy. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Hiroshi Echizen'ya, Kenji Araki, and Eduard Hovy. 2019. Word embedding-based automatic MT evaluation metric using word position information. 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 1874–1883, Minneapolis, Minnesota. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020a. Bergamot-latte submissions for the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1008–1015, Online. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020b. Unsupervised quality estimation for neural machine translation. *TACL*, 8:539–555.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Alon Lavie and Abhaya Agarwal. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. pages 65–72.

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

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

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

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2019. Vifidel: Evaluating the visual fidelity of image descriptions.

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

Nitika Mathur, Johnny Wei, Qingsong Ma, and Ondrej Bojar. 2020. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 686–723, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3:1–12.

João Moura, miguel vera, Daan van Stigt, Fabio Kepler, and Andre F. T. Martins. 2020. Ist-unbabel participation in the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1027–1034, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

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 1047–1053, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Kashif Shah, Trevor Cohn, and Lucia Specia. 2015. A bayesian non-linear method for feature selection in machine translation quality estimation. In *Machine Translation 29*, pages 79–84.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and Andre F. T. Martins. 2020. Findings of the wmt 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 741–762, Online. Association for Computational Linguistics.

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

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020b. BERT-XML: Large scale automated ICD coding using BERT pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 24–34, 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. *CoRR*, abs/1909.02622.

A Appendix

A.1 Layers of Embeddings

Since both XLM-Roberta-Base and Roberta-Large have multiple layers, selecting a good layer or combination of layers is important for WMD and BERTScore. Therefore, we use WMT-17 as our development dataset to study our representation choices. The Pearson correlation of WMD with human judgement scores SRC-MT by choosing specific XLM-Roberta-Base's layers is shown in Figure 3. Selecting Layer 9 as the token embeddings for XLM-Roberta-Base leads to the best average Pearson correlation among 9 language pairs for the SRC-MT setting

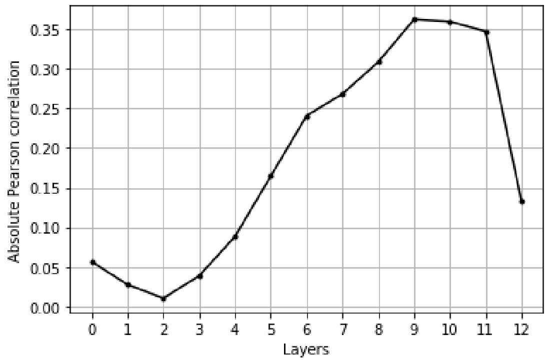

Figure 3: Pearson correlation of WMD with different layers of XLM-Roberta-Base embeddings.

For Roberta-Large, we show in Figure 4 the use of WMT-17 MT-REF task to evaluate the performance of different layer. Among the 24 output layers, the recommended layer of representation to use is 17. This is inline with the results described in (Zhang et al., 2020a), where the best layer for Roberta-Large to use in BERTScore is also found to be Layer 17.

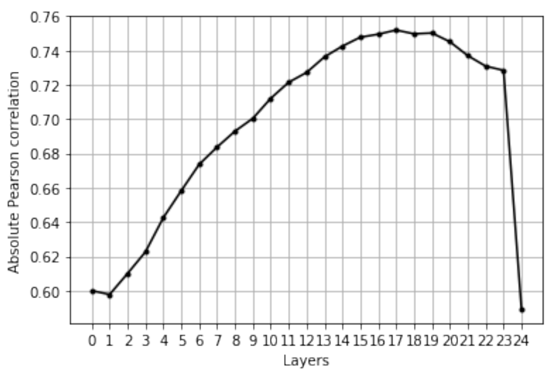

Figure 4: Pearson correlation of WMD with different layers of Roberta-Large embeddings.

A.2 Performance of Sentence Mover's Distance (SMD)

Sentence Mover's Distance (SMD) (Zhao et al., 2019) is an alternative metric which for sentence semantic similarity. It compares two text documents using sentence embeddings which are not fine-tuned but based on averaging or pooling the sentences' combined contextual word embeddings. The SMD is defined as follows:

$$SMD(x^n, y^n) := \|E(x_1^{l_x}) - E(y_1^{l_y})\| \quad (8)$$

where E is the embedding function which maps an n-gram to its vector representation, l_x and l_y are the size of sentences. As a comparison, we experimented with the linear combination between SMD and each or our token-level metrics – WMD and BERTScore. The metrics performances for WMT-17 in both cases of SRC-MT and MT-REF, and WMT-20 SRC-MT are shown in Table 9, Table 8 and Table 7.

The overall performance of this metric is inferior to that of SSS, which is to be expected since this is simply averaging token-level embeddings. Similar to our SSS, the SMD metric performance improve when combined with token-level metrics. The combined metrics' performance drops when there is a big gap on the scores of the two combined metrics, such as more than 10%. For example, the results in Tables 9 and 7 show the following behaviour:

1. **zh-en**: Performance gap is 11.5%, **cs-en**: Performance gap is 16.1%, **en-ru**: Performance gap is 16.3%, **tr-en**: Performance gap is 13.1%. All performance gaps are computed based on the difference between SMD and BERTScore.

2. **en-de**: Performance gap is 9.1%, **et-en**: Performance gap is 15.7%, **ro-en**: performance gap is 10.2%. All performance gaps are computed based on the difference between SMD and BERTScore.

A.3 Plots with Metrics' Performance

To facilitate visualisation of our main tabular results presented in the paper, Figures 5, 6, 7 graphically summarize the performance of SRC-MT and MT-REF cases for the WMT-17 and WMT-20 datasets. In all cases, with the addition of SSS (Similarity), both BERTScore and WMD have improved performance, which is also better than SSS

Metrics	ne-en	en-de	et-en	en-zh	ro-en	si-en	ru-en	Avg
WMD	0.361	0.456	0.463	0.251	0.647	0.308	0.315	0.400
SMD	0.436	0.368	0.302	0.277	0.570	0.298	0.281	0.362
BERTScore	0.357	**0.459**	**0.460**	0.260	**0.673**	0.309	0.320	0.405
SMD + WMD	**0.452**	0.423	0.401	0.279	0.618	0.355	0.326	0.408
SMD + BERTScore	0.449	0.439	0.413	**0.289**	0.638	**0.363**	**0.327**	**0.417**

Table 7: Pearson correlation with human scores in WMT-20 SRC-MT case with Sentence Mover's Distance (SMD).

Metrics	de-en	zh-en	fi-en	lv-en	ru-en	cs-en	tr-en	Avg
WMD	0.730	0.769	0.827	0.736	0.733	0.698	0.770	0.752
SMD	0.703	0.686	0.763	0.693	0.698	0.648	0.644	0.691
BERTScore	0.745	**0.775**	0.833	0.756	0.746	0.710	0.751	0.759
SMD + WMD	0.745	0.757	0.832	0.750	0.736	0.705	**0.753**	0.754
SMD + BERTScore	**0.757**	0.771	**0.846**	**0.764**	**0.752**	**0.717**	0.752	**0.766**

Table 8: Pearson correlation with human scores in WMT-17 MT-REF case with Sentence Mover's Distance (SMD).

Metrics	de-en	zh-en	fi-en	lv-en	ru-en	cs-en	en-ru	en-zh	tr-en	Avg
WMD	0.366	0.501	0.373	0.373	0.308	0.267	0.404	0.408	0.350	0.372
SMD	0.348	0.394	0.360	0.342	0.276	0.158	0.271	0.345	0.250	0.305
BERTScore	0.409	**0.510**	0.414	**0.402**	0.337	**0.319**	**0.434**	0.446	**0.382**	**0.406**
SMD + WMD	0.392	0.491	0.392	0.382	0.343	0.239	0.373	0.429	0.310	0.372
SMD + BERTScore	**0.417**	0.503	**0.416**	0.400	**0.361**	0.271	0.394	**0.454**	0.341	0.395

Table 9: Pearson correlation with human scores in WMT-17 SRC-MT case with Sentence Mover's Distance (SMD).

alone (red lines on bars), except for ru-en SRC-MT case in WMT-17, where the performance of SSS + WMD is lower than that of SSS.

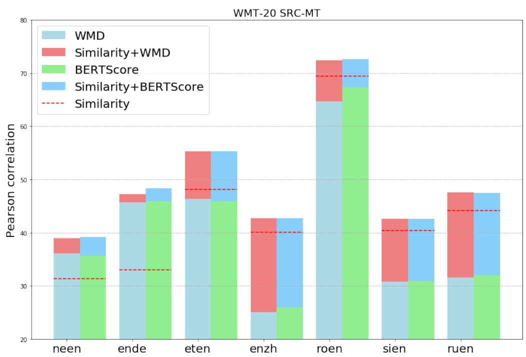

Figure 5: Metrics' performance in WMT-20 SRC-MT case.

A.4 Performance with Roberta-Base embeddings

We present performance of Roberta-Base in the case of WMT-17 MT-REF in Table 10. This shows that stronger embeddings (XLM-Roberta-Base, Table5) indeed will lead to better performance.

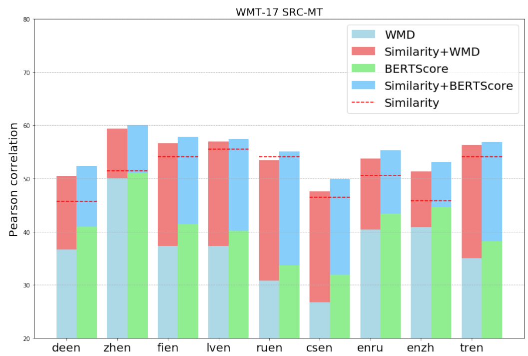

Figure 6: Metrics' performance in WMT-17 SRC-MT case.

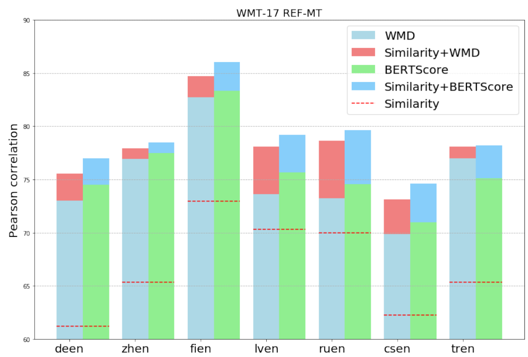

Figure 7: Metrics' performance in WMT-17 MT-REF case.

Metrics	de-en	zh-en	fi-en	lv-en	ru-en	cs-en	tr-en	Avg
WMD	0.667	0.743	0.818	0.693	0.705	0.663	0.744	0.719
SSS	0.612	0.655	0.705	0.680	0.642	0.599	0.644	0.648
BERTScore	0.683	0.740	0.818	0.693	0.707	0.675	0.718	0.719
SSS + WMD	0.718	0.767	0.832	0.755	0.736	0.703	**0.764**	0.754
SSS + BERTScore	**0.728**	**0.767**	**0.843**	**0.755**	**0.744**	**0.717**	0.758	**0.759**

Table 10: Pearson correlation with human score of WMT-17 with Roberta-Base in the case of MT-REF (To English).