A Appendix

A.1 Comparison to Sentence Mover's Distance

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

$$SMD(x^n, y^n) := \|E(x_1^{l_x}) - E(y_1^{l_y})\| \quad (9)$$

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 of 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 8, Table 9 and Table 10.

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 improves when it is combined with token-level metrics. The combined metrics' performance drops when there is a big difference between the scores of the two combined metrics, e.g. more than 10%. To pick an example, in Table 8 the gap between BERTScore and SMD for **zh-en** is 0.115, and the combined SMD + BERTScore only reaches a score of 0.503, compared to 0.51 from BERTScore alone. For other languages with closer BERTScore and SMD scores, the performance of the combined metric remains the same or improves, for example, **ru-en**.

A.2 Plots with Metrics' Performance

To facilitate visualisation of our main tabular results presented in the paper, Figures 5, 6, 7 show them as bar plots.

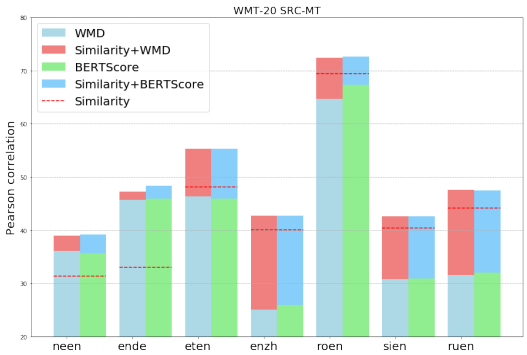

Figure 5: Metrics' performance in WMT-20 SRC-MT case.

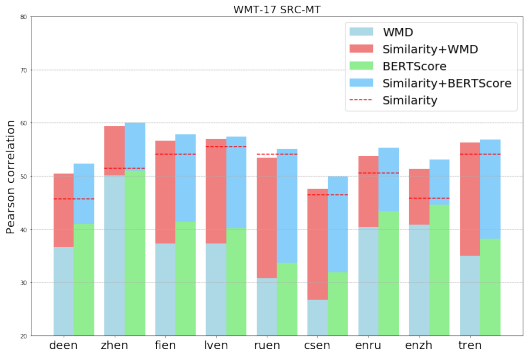

Figure 6: Metrics' performance in WMT-17 SRC-MT case.

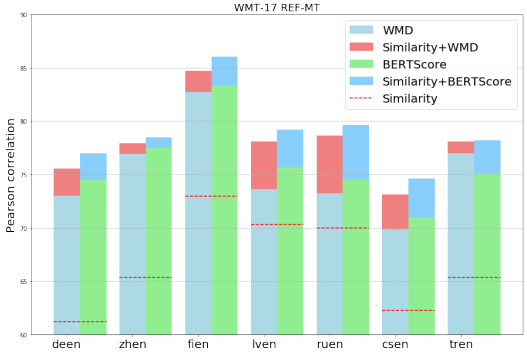

Figure 7: Metrics' performance in WMT-17 MT-REF case.

Metrics	de-en	zh-en	fi-en	lv-en	ru-en	cs-en	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
BERTScore	0.409	**0.510**	0.414	0.**402**	0.337	**0.319**	**0.434**	0.446	**0.382**	**0.406**
SMD	0.348	0.394	0.360	0.342	0.276	0.158	0.271	0.345	0.250	0.305
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 8: Pearson Correlation with human scores in WMT-17 SRC-MT case with Sentence Mover's Distance.

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
BERTScore	0.745	**0.775**	0.833	0.756	0.746	0.710	0.751	0.759
SMD	0.703	0.686	0.763	0.693	0.698	0.648	0.644	0.691
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 9: Pearson Correlation with human scores in WMT-17 MT-REF case with Sentence Mover's Distance.

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
BERTScore	0.357	**0.459**	**0.460**	0.260	**0.673**	0.309	0.320	0.405
SMD	0.436	0.368	0.302	0.277	0.570	0.298	0.281	0.362
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 10: Pearson Correlation with human scores in WMT-20 SRC-MT case with Sentence Mover's Distance.