A Consistent advantage of batch centering with different layers

(Zhang et al., 2019b) did an extensive search of the best layer was done for models using WMT-16 dataset. While the best layer varies from model to model in general, the best layers for base and large versions of BERT and RoBERTa are found to be very close to the 10th and the 19th.

We conduct the experiments for STS 12-16 and WMT-17 for five different layers and show the result in Fig. 5 and Fig. 6. The batch centered version (solid lines) of all the metrics perform consistently better than its uncentered counterparts (dashed lines). For WMT-18, we selected two more layers for roberta-base and roberta large (Table 5), where we also see the consistent advantage of batch centering. The values of Pearson correlation have little change from layer to layer as well.

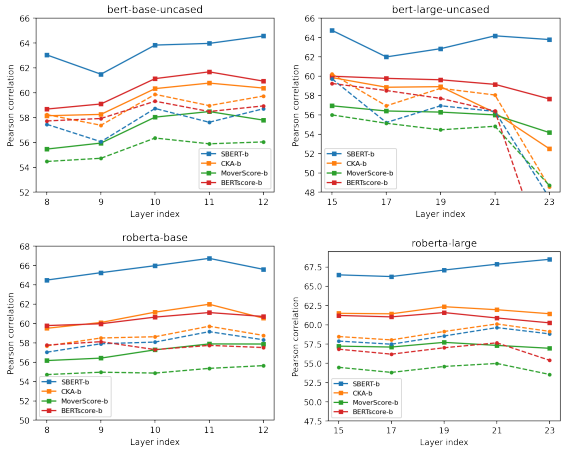

Figure 5: Four models with four metrics evaluated on STS 12-16 using different layers. Squares connected by solid lines are with batch centering. Dots connected by dashed lines are their counterparts without batch centering.

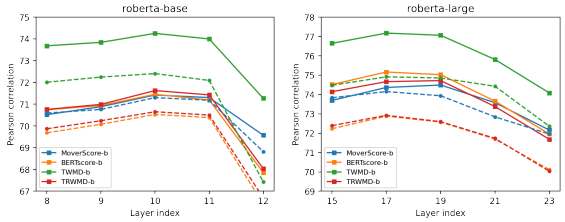

Figure 6: roberta-base and roberta-large with four metrics evaluated on WMT-17 using different layers. Squares connected by solid lines are with batch centering. Dots connected by dashed lines are their counterparts without batch centering.

Table 5: WMT-18 evaluated with two more layers for roberta-base and large.

Metric	**Avg.** τ / r	
roberta-base	**Layer-9**	**Layer-11**
SBERT	26.2 / 36.0	27.1 / 37.0
SBERT-b	29.6 / 41.7	29.3 / 41.3
CKA	26.3 / 36.3	27.1 / 37.4
CKA-b	30.1 / 43.2	30.0 / 43.3
MoverScore	30.4 / 42.5	30.5 / 42.8
MoverScore-b	30.5 / 42.5	30.5 / 42.7
BERTscore	30.0 / 42.7	30.2 / 42.9
BERTscore-b	30.2 / 43.0	30.2 / 43.1
TWMD	30.9 / 43.6	30.9 / 43.5
TWMD-b	31.5 / 44.3	31.4 / 44.4
TRWMD	30.1 / 42.7	30.3 / 42.8
TRWMD-b	30.5 / 42.9	30.6 / 43.1
roberta-large	**Layer-17**	**Layer-21**
SBERT	29.6 / 40.5	28.7 / 40.4
SBERT-b	31.4 / 43.7	30.4 / 42.5
CKA	29.6 / 40.9	28.8 / 40.7
CKA-b	31.4 / 45.0	30.8 / 44.3
MoverScore	31.8 / 43.9	31.3 / 43.1
MoverScore-b	31.6 / 43.7	31.3 / 43.3
BERTscore	31.8 / 44.1	30.8 / 43.2
BERTscore-b	31.6 / 44.8	31.0 / 44.0
TWMD	32.5 / 44.9	31.8 / 44.5
TWMD-b	32.7 / 45.9	32.1 / 45.1
TRWMD	31.7 / 44.0	30.8 / 43.2
TRWMD-b	31.8 / 44.4	31.2 / 43.7

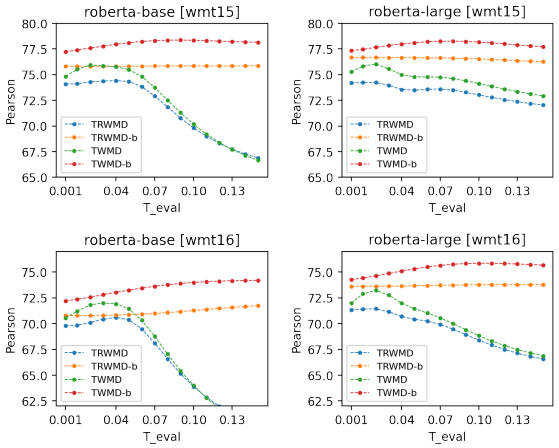

Figure 7: Pearson correlation vs. temperature in evaluation of WMT 15-16.

B Choice of temperature using WMT-15 and 16 as validation

We use the datasets WMT-15 and WMT-16 to determine the optimal temperature for TWMD, TRWMD, TWMD-b and TRWMD-b, then fix the temperature for evaluating WMT-17 (Table 2) and WMT-18 (Table 3).

In particular, we first estimate the Pearson cor-

relation for each metric with roberta-base and roberta-large on WMT-15 and 16. Next we compute the weighted mean of WMT-15 and 16 by $\bar{r} = (N_{15} * r_{15} + N_{16} * r_{16})/(N_{15} + N_{16})$ where N and r is the size of dataset and the Pearson correlation. Last, we take the mean of $\bar{r}$ for roberta-base and roberta-large and use the result $r = (\bar{r}_{\text{base}} + \bar{r}_{\text{large}})/2$ to determine the optimal temperature for each metric. We thus obtained the best temperature for TWMD and TRWMD to be 0.02, and the best temperatures for TWMD-b and TRWMD-b to be 0.1 and 0.15, respectively.

C Precision and F1 Scores

The BERTscore-Recall (Eq. 3) is an asymmetic metric. (Zhang et al., 2019b) also provides two additional metrics: the BERTscore-Precision that switches the roles of the query and reference sentences; and a symmetric BERTscore-F1 metric that is the harmonic mean of BERTscore-Recall and BERTscore-Precision metrics. Since BERTscore-F1 is an ensemble of the two metrics, it is expected to perform better. However, according to (Zhang et al., 2019b), BERTscore-Precision and BERTscore-F1 are inconsistent and sometimes significantly underperform (e.g. in COCO image captioning).

Similar to BERTscore, both TRWMD and TWMD metrics are asymmetric[‡]. In Table 6 and 7, we present the results of the precision-scores of BERTscore (Zhang et al., 2019b), TRWMD and TWMD with or without batch centering in WMT-17 and 18, where the temperatures are the same as those for the recall-scores (see Appendix B). In Table 8 and 9, we present the results of the F1-scores, where the temperature for TWMD, TRWMD, TWMD-b and TRWMD-b are 0.01, 0.01, 0.08 and 0.06 respectively. TWMD-b is the top performer in most of cases, albeit the lead shrinks in the cases of the F1-scores. In addition, batch-centering produces consistent improvements for all metrics.

[‡]TWMD is asymmetric when only few Sinkhorn iterations are applied.

Table 6: Correlation with human scores on the WMT-17 Metrics Shared Task using the Precision scores for BERTscore, TWMD and TRWMD with or without batch centering.

Metric	cs-en τ / r	de-en τ / r	fi-en τ / r	lv-en τ / r	ru-en τ / r	tr-en τ / r	zh-en τ / r	Avg. τ / r
roberta-base								
BERTscore	47.9 / 66.1	47.5 / 66.0	56.9 / 76.9	48.0 / 67.1	50.7 / 68.6	51.3 / 70.3	53.5 / 74.6	50.8 / 70.0
BERTscore-b	48.6 / 67.8	48.5 / 67.7	58.5 / 78.7	48.1 / 68.5	51.1 / 70.3	53.8 / 73.1	52.3 / 73.8	51.6 / 71.3
TWMD	48.8 / 66.4	49.2 / 67.9	61.9 / 80.9	51.0 / 70.3	52.6 / 71.8	54.0 / 73.6	**55.0 / 75.9**	53.2 / 72.3
TWMD-b	**50.8 / 69.5**	**51.7 / 70.8**	**63.4 / 83.1**	**52.4 / 72.5**	**54.0 / 73.9**	**56.8 / 77.1**	54.3 / 75.4	**54.8 / 74.6**
TRWMD	47.9 / 65.9	47.6 / 66.5	57.0 / 77.0	48.0 / 67.0	50.8 / 68.8	51.4 / 71.1	53.4 / 74.5	50.9 / 70.1
TRWMD-b	49.1 / 67.7	49.7 / 68.5	58.6 / 79.3	48.5 / 68.2	52.0 / 70.7	54.3 / 74.6	52.4 / 73.5	52.1 / 71.8
roberta-large								
BERTscore	51.9 / 69.5	54.3 / 72.8	57.3 / 77.2	51.3 / 70.1	54.6 / 71.7	53.1 / 72.5	56.2 / 76.1	54.1 / 72.8
BERTscore-b	52.2 / 72.0	54.3 / 73.5	60.0 / 80.0	52.2 / 72.8	55.4 / 74.4	56.2 / 75.4	56.0 / 76.7	55.2 / 75.0
TWMD	52.2 / 69.0	55.5 / 74.0	62.7 / 81.2	54.0 / 72.8	56.0 / 74.4	56.1 / 74.7	**57.9 / 78.0**	56.3 / 74.9
TWMD-b	**54.6 / 73.9**	**56.8 / 76.0**	**64.5 / 83.6**	**55.7 / 75.5**	**57.4 / 76.5**	**58.1 / 78.3**	57.4 / **78.0**	**57.8 / 77.4**
TRWMD	51.5 / 68.8	54.3 / 72.7	56.8 / 76.9	51.0 / 69.8	54.5 / 71.7	52.8 / 72.5	56.2 / 76.0	53.9 / 72.6
TRWMD-b	52.4 / 71.3	54.7 / 73.5	59.4 / 79.7	52.0 / 71.8	55.3 / 73.4	55.7 / 76.0	55.6 / 76.0	55.0 / 74.5

Table 7: Correlation with human scores on the WMT-18 Metrics Shared Task using the Precision scores for BERTscore, TWMD and TRWMD with or without batch centering.

Metric	cs-en τ / r	zh-en τ / r	ru-en τ / r	fi-en τ / r	tr-en τ / r	et-en τ / r	de-en τ / r	Avg. τ / r
roberta-base								
BERTscore	28.9 / 40.4	27.7 / 37.8	27.3 / 39.4	25.2 / 36.5	30.8 / 43.1	33.9 / 48.1	39.1 / 55.0	30.4 / 42.9
BERTscore-b	29.2 / 41.6	27.5 / 38.1	27.5 / 39.9	25.4 / 37.2	30.9 / 43.2	34.2 / 48.9	39.4 / 55.6	30.5 / 43.5
TWMD	29.2 / 41.3	28.4 / 38.4	28.1 / 40.2	25.6 / 36.7	31.5 / 43.6	34.8 / 49.3	39.9 / 56.4	31.1 / 43.7
TWMD-b	**29.7 / 42.3**	**28.7 / 39.1**	**28.7 / 40.9**	**26.4 / 38.3**	**32.2 / 44.4**	**35.4 / 50.3**	**40.4 / 57.3**	**31.6 / 44.7**
TRWMD	28.9 / 40.4	27.7 / 37.8	27.4 / 39.4	25.1 / 36.1	30.9 / 43.0	33.9 / 48.1	39.2 / 55.1	30.4 / 42.9
TRWMD-b	29.4 / 41.6	27.9 / 37.9	28.1 / 39.9	25.3 / 36.6	31.2 / 43.0	34.5 / 49.1	39.8 / 55.9	30.9 / 43.4
roberta-large								
BERTscore	30.0 / 41.6	28.2 / 38.3	29.0 / 40.9	26.7 / 37.8	31.6 / 43.6	35.9 / 49.8	41.0 / 57.4	31.8 / 44.2
BERTscore-b	30.5 / 43.7	28.1 / 38.7	28.9 / 41.3	27.1 / 39.1	31.5 / 44.2	35.9 / 50.6	41.0 / 57.5	31.9 / 45.0
TWMD	30.7 / 42.9	28.9 / 39.1	29.5 / 41.1	27.3 / 38.5	32.3 / 44.3	36.6 / 50.7	41.8 / 58.6	32.5 / 45.1
TWMD-b	**31.3 / 44.4**	**29.0 / 39.5**	**29.9 / 41.9**	**27.7 / 39.8**	**32.6 / 45.0**	**36.8 / 51.5**	**41.9 / 59.0**	**32.8 / 45.9**
TRWMD	29.7 / 41.3	28.2 / 38.2	29.0 / 40.6	26.5 / 37.4	31.6 / 43.4	35.8 / 49.7	41.0 / 57.3	31.7 / 44.0
TRWMD-b	30.4 / 43.1	28.2 / 38.3	29.2 / 41.0	26.7 / 38.2	31.7 / 43.7	35.9 / 50.4	41.1 / 57.5	31.9 / 44.6

Table 8: Correlation with human scores on the WMT-17 Metrics Shared Task using the F1 scores for BERTscore, TWMD and TRWMD with or without batch centering.

Metric	cs-en τ / r	de-en τ / r	fi-en τ / r	lv-en τ / r	ru-en τ / r	tr-en τ / r	zh-en τ / r	Avg. τ / r
roberta-base								
BERTscore	**50.2** / 68.8	50.3 / 69.3	62.9 / 82.0	51.3 / 71.1	53.0 / 72.1	54.6 / 74.0	54.4 / 75.5	53.8 / 73.3
BERTscore-b	**50.2 / 69.3**	51.0 / 70.5	62.9 / 82.3	50.9 / 71.8	53.0 / 72.8	55.4 / 75.2	52.9 / 74.4	53.8 / 73.8
TWMD	48.4 / 66.3	49.5 / 68.5	62.3 / 81.0	51.5 / 70.9	52.6 / 72.0	54.5 / 73.7	**55.2 / 76.0**	53.4 / 72.6
TWMD-b	50.0 / 68.8	**51.4 / 70.8**	**63.1 / 82.9**	**52.1 / 72.5**	**53.6 / 73.6**	**56.5 / 76.6**	54.1 / 75.4	**54.4 / 74.4**
TRWMD	**50.2** / 68.8	50.4 / 69.3	63.0 / 82.0	51.4 / 71.1	53.1 / 72.1	54.7 / 74.1	54.5 / 75.6	53.9 / 73.3
TRWMD-b	**50.2 / 69.3**	51.1 / 70.6	63.0 / 82.3	51.0 / 71.8	53.2 / 73.1	55.6 / 75.4	53.3 / 74.6	53.9 / 73.9
roberta-large								
BERTscore	54.0 / 72.0	56.2 / 75.0	63.1 / 81.8	54.8 / 73.5	56.2 / 73.9	56.3 / 75.4	57.4 / 77.6	56.8 / 75.6
BERTscore-b	54.1 / **74.0**	56.2 / 75.9	**64.6** / 83.5	55.3 / 75.9	56.8 / 76.2	57.4 / 77.0	56.4 / 77.3	57.3 / 77.1
TWMD	52.5 / 69.3	55.8 / 74.5	63.5 / 81.5	54.6 / 73.5	56.2 / 74.6	56.6 / 74.8	**58.0 / 78.2**	56.7 / 75.2
TWMD-b	54.1 / 73.5	56.2 / 75.8	**64.6 / 83.6**	55.4 / 75.6	**57.2 / 76.7**	**58.0 / 77.8**	57.1 / 77.8	**57.5 / 77.3**
TRWMD	54.0 / 72.1	**56.3** / 75.1	63.1 / 81.8	54.8 / 73.5	56.3 / 74.0	56.3 / 75.4	57.5 / 77.7	56.8 / 75.7
TRWMD-b	**54.3 / 74.0**	**56.3 / 76.0**	**64.6** / 83.5	**55.5 / 76.0**	56.9 / 76.4	57.5 / 77.2	56.8 / 77.5	57.4 / 77.2

Table 9: Correlation with human scores on the WMT-18 Metrics Shared Task using the F1 scores for BERTscore, TWMD and TRWMD with or without batch centering.

Metric	cs-en τ / r	zh-en τ / r	ru-en τ / r	fi-en τ / r	tr-en τ / r	et-en τ / r	de-en τ / r	Avg. τ / r
roberta-base								
BERTscore	**29.5** / 41.9	28.4 / 39.1	28.4 / **41.0**	26.1 / 37.9	31.9 / **44.6**	35.0 / 49.5	40.2 / 56.9	31.4 / 44.4
BERTscore-b	**29.5 / 42.1**	28.1 / 39.1	28.4 / **41.0**	**26.4 / 38.6**	31.9 / **44.6**	35.1 / 49.9	40.3 / 57.0	31.4 / 44.6
TWMD	29.0 / 41.1	28.4 / 38.5	28.1 / 40.2	25.6 / 36.8	31.7 / 43.9	34.9 / 49.4	40.0 / 56.6	31.1 / 43.8
TWMD-b	**29.5 / 42.1**	**28.5** / 39.0	**28.7** / 40.8	**26.4** / 38.3	**32.2** / 44.5	**35.5 / 50.3**	**40.4 / 57.4**	**31.7 / 44.7**
TRWMD	**29.5** / 41.8	28.4 / 39.1	28.5 / **41.0**	26.1 / 37.8	31.9 / **44.6**	35.0 / 49.5	40.2 / 56.9	31.4 / 44.4
TRWMD-b	**29.5 / 42.1**	28.4 / **39.2**	28.4 / 40.9	**26.4 / 38.6**	32.0 / **44.6**	35.1 / 49.9	40.3 / 57.1	31.5 / 44.6
roberta-large								
BERTscore	30.8 / 43.3	28.9 / 39.3	**29.9** / 41.6	27.4 / 38.7	32.4 / 44.7	36.7 / 50.6	**42.0 / 59.0**	32.6 / 45.3
BERTscore-b	31.2 / **44.6**	28.9 / **39.8**	29.7 / **42.4**	**27.8 / 40.3**	**32.5** / 45.4	**36.8** / 51.2	41.8 / **59.0**	**32.7 / 46.1**
TWMD-b	30.8 / 43.2	**29.0** / 39.4	29.5 / 41.3	27.5 / 38.8	**32.5** / 44.7	36.7 / 50.7	41.8 / 58.9	32.6 / 45.3
TWMD-b	**31.3** / 44.5	**29.0** / 39.7	29.7 / 42.0	27.7 / 39.9	**32.5** / 45.2	**36.8 / 51.4**	41.8 / **59.0**	**32.7** / 46.0
TRWMD	30.8 / 43.3	**29.0** / 39.3	**29.9** / 41.6	27.4 / 38.7	32.4 / 44.7	36.7 / 50.6	**42.0 / 59.0**	32.6 / 45.3
TRWMD-b	31.2 / **44.6**	28.9 / **39.8**	**29.9 / 42.4**	**27.8 / 40.3**	**32.5 / 45.5**	**36.8** / 51.1	41.8 / **59.0**	**32.7 / 46.1**