A Supplemental Material

A.1 Proofs

In this section, we prove the Prop. 1 in the paper about viewing BERTScore (precision/recall) as a (non-optimized) Mover Distance.

As a reminder, WMD formulates as:

$$\text{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{F \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \sum_i \sum_j C_{ij} \cdot F_{ij}$$
$$\text{s.t. } \mathbf{1}^\intercal F^\intercal \mathbf{1} = 1, \quad \mathbf{1}^\intercal F \mathbf{1} = 1.$$

where $F^\intercal \mathbf{1} = \boldsymbol{f}_x^n$ and $F\mathbf{1} = \boldsymbol{f}_y^n$ denote a vector of weights for each n-gram of $\boldsymbol{x}^n$ and $\boldsymbol{y}^n$.

The BERTScore is defined as:

$$R_{\text{BERT}} = \frac{\sum_{y_i^1 \in \boldsymbol{y}^1} \text{idf}(y_i^1) \max_{x_j^1 \in \boldsymbol{x}^1} E(x_j^1)^\intercal E(y_i^1)}{\sum_{y_i^1 \in \boldsymbol{y}^1} \text{idf}(y_i^1)}$$
$$P_{\text{BERT}} = \frac{\sum_{x_j^1 \in \boldsymbol{x}^1} \text{idf}(x_j^1) \max_{y_i^1 \in \boldsymbol{y}^1} E(y_i^1)^\intercal E(x_j^1)}{\sum_{x_j^1 \in \boldsymbol{x}^1} \text{idf}(x_j^1)}$$
$$F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad .$$

Then, R_{BERT} can be formulated in a "quasi" WMD form:

$$R_{\text{BERT}}(\boldsymbol{x}^1, \boldsymbol{y}^1) := \sum_{i,j} C_{ij} \cdot F_{ij}$$
$$F_{ij} = \begin{cases} \frac{1}{M} & \text{if } x_j = \arg\max_{\hat{x}_j^1 \in \boldsymbol{x}^1} E(y_i^1)^\intercal E(\hat{x}_j^1) \\ 0 & \text{otherwise} \end{cases}$$
$$C_{ij} = \begin{cases} \frac{M}{Z} \text{idf}(y_i^1) E(x_j^1)^\intercal E(y_i^1) & \text{if } x_j = \arg\max_{\hat{x}_j^1 \in \boldsymbol{x}^1} E(y_i^1)^\intercal E(\hat{x}_j^1) \\ 0 & \text{otherwise} \end{cases}$$

where $Z = \sum_{y_i^1 \in \boldsymbol{y}^1} \text{idf}(y_i^1)$ and M is the size of n-grams in $\boldsymbol{x}^1$. Similarly, we can have P_{BERT} in a quasi WMD form (ignored). Then, F_{BERT} can be formulated as harmonic-mean of two WMD forms of P_{BERT} and F_{BERT}.

A.2 Routing

In this section, we study the aggregation function ϕ with routing scheme, which achieved promising results on NLP tasks. Specifically, we introduce a nonparametric clustering with Kernel Density Estimation (KDE) for routing since KDE bridges a family of kernel functions with underlying empirical distributions, which often leads to computational efficiency (Zhang et al., 2018), defined as:

$$\min_{\boldsymbol{v}, \gamma} f(\boldsymbol{z}) = \sum_{i=1}^{L} \sum_{j=1}^{T} \gamma_{ij} k(d(\boldsymbol{v}_j - \boldsymbol{z}_{i,j}))$$
$$s.t. \quad \forall i, j : \gamma_{ij} > 0, \sum_{j=1}^{L} \gamma_{ij} = 1.$$

where $d(.)$ is a distance function, γ_{ij} denotes underlying closeness between the aggregated vector $\boldsymbol{v}_j$ and vector $\boldsymbol{z}_i$ in the i-th layer, some instances of $k(.)$ (Wand and Jones, 1994) can be illustrated as:

$$Gaussian : k(x) \triangleq \exp\left(-\frac{x}{2}\right), \quad Epanechnikov : k(x) \triangleq \begin{cases} 1 - x & x \in [0, 1) \\ 0 & x \geq 1. \end{cases}$$

One typical solution for KDE clustering to minimize $f(\boldsymbol{z})$ is taking Mean Shift (Comaniciu and Meer, 2002), defined as:

$$\nabla f(\boldsymbol{z}) = \sum_{i,j} \gamma_{ij} k'(d(\boldsymbol{v}_j, \boldsymbol{z}_{i,j})) \frac{\partial d(\boldsymbol{v}_j, \boldsymbol{z}_{i,j})}{\partial \boldsymbol{v}}$$

Firstly, $\boldsymbol{v}_j^{\tau+1}$ can be updated while $c_{ij}^{\tau+1}$ is fixed:

$$\boldsymbol{v}_j^{\tau+1} = \frac{\sum_i \gamma_{ij}^{\tau} k'(d(\boldsymbol{v}_j^{\tau}, \boldsymbol{z}_{i,j}))\boldsymbol{z}_{i,j}}{\sum_{i,j} k'(d(\boldsymbol{v}_j^{\tau}, \boldsymbol{z}_{i,j}))}$$

Intuitively, $\boldsymbol{v}_j$ can be explained as a final aggregated vector from L contextualized layers. Then, we adopt SGD to update $\gamma_{ij}^{\tau+1}$:

$$\gamma_{ij}^{\tau+1} = \gamma_{ij}^{\tau} + \alpha \cdot k(d(\boldsymbol{v}_j^{\tau}, \boldsymbol{z}_{i,j}))$$

where α is the hyper parameter to control step size. The routing process is summarized in Algorithm 1.

Algorithm 1 Aggregation by Routing

procedure ROUTING($\boldsymbol{z}_{ij}$, ℓ)
Initialize $\forall i, j : \gamma_{ij} = 0$
while true **do**
  foreach representation i and j in layer ℓ and $\ell + 1$ **do** $\gamma_{ij} \leftarrow softmax(\gamma_{ij})$
  foreach representation j in layer $\ell + 1$ **do**
    $\boldsymbol{v}_j \leftarrow \sum_i \gamma_{ij} k'(\boldsymbol{v}_j, \boldsymbol{z}_i)\boldsymbol{z}_i / \sum_i k'(\boldsymbol{v}_i, \boldsymbol{z}_i)$
  foreach representation i and j in layer ℓ and $\ell + 1$ **do** $\gamma_{ij} \leftarrow \gamma_{ij} + \alpha \cdot k(\boldsymbol{v}_j, \boldsymbol{z}_i)$
  loss $\leftarrow \log(\sum_{\mathrm{i,j}} \gamma_{\mathrm{ij}} \mathrm{k}(\boldsymbol{v}_{\mathrm{j}}, \boldsymbol{z}_{\mathrm{i}}))$
  if $|\text{loss} - \text{preloss}| < \epsilon$ **then**
    break
  else
    preloss $\leftarrow$ loss
return v_j

Best Layer and Layer-wise Consolidation Table 6 compares our Word Mover based metric combining BERT representations on different layers with stronger BERT representations consolidated from these layers (p-means and routing). We often see that which layer has best performance is task-dependent, and our WMD-based metrics with p-means or routing schema come close to the oracle performance obtained from best layers.

	Direct Assessment						
Metrics	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en
WMD-1 + BERT + LAYER 8	.6361	.6755	.8134	.7033	.7273	.7233	.7175
WMD-1 + BERT + LAYER 9	.6510	.6865	.8240	.7107	.7291	**.7357**	**.7195**
WMD-1 + BERT + LAYER 10	.6605	**.6948**	**.8231**	.7158	**.7363**	.7317	.7168
WMD-1 + BERT + LAYER 11	**.6695**	.6845	.8192	.7048	.7315	.7276	.7058
WMD-1 + BERT + LAYER 12	.6677	.6825	.8194	**.7188**	.7326	.7291	.7064
WMD-1 + BERT + ROUTING	.6618	.6897	.8225	.7122	.7334	.7301	.7182
WMD-1 + BERT + PMEANS	.6623	.6873	.8234	.7139	.7350	.7339	.7192

Table 6: Absolute Pearson correlations with segment-level human judgments on WMT17 to-English translations.

Experiments Table 7, 8 and 9 show correlations between metrics (all baseline metrics and Word Mover based metrics) and human judgments on machine translation, text summarization and dialogue response generation tasks.

Setting	Metrics	Direct Assessment cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
Existing Metrics	BLEND	0.594	0.571	0.733	0.594	0.622	0.671	0.661	0.635
	RUSE	0.624	0.644	0.750	0.697	0.673	0.716	0.691	0.685
	SENTBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.481
	CHRF++	0.523	0.534	0.678	0.520	0.588	0.614	0.593	0.579
	METEOR++	0.552	0.538	0.720	0.563	0.627	0.626	0.646	0.610
	BERTSCORE-F1	0.670	0.686	0.820	0.710	0.729	0.714	0.704	0.719
Word-Mover	WMD-1 + W2V	0.392	0.463	0.558	0.463	0.456	0.485	0.481	0.471
	WMD-1 + BERT + ROUTING	0.658	0.689	0.823	0.712	0.733	0.730	0.718	0.723
	WMD-1 + BERT + MNLI + ROUTING	0.665	0.705	0.834	0.744	0.735	0.752	0.736	0.739
	WMD-2 + BERT + MNLI + ROUTING	0.676	0.706	0.831	0.743	0.734	0.755	0.732	0.740
	WMD-1 + BERT + PMEANS	0.662	0.687	0.823	0.714	0.735	0.734	0.719	0.725
	WMD-1 + BERT + MNLI + PMEANS	0.670	0.708	0.835	0.746	0.738	0.762	0.744	0.743
	WMD-2 + BERT + MNLI + PMEANS	0.679	0.710	0.832	0.745	0.736	0.763	0.740	0.743

Table 7: Absolute Pearson correlations with segment-level human judgments on WMT17 to-English translations.

Setting	Metrics	TAC-2008 Responsiveness r	Responsiveness ρ	Pyramid r	Pyramid ρ	TAC-2009 Responsiveness r	Responsiveness ρ	Pyramid r	Pyramid ρ
Existing Metrics	S^3_{full}	0.696	0.558	0.753	0.652	0.731	0.552	0.838	0.724
	S^3_{best}	0.715	0.595	0.754	0.652	0.738	0.595	0.842	0.731
	TF∗IDF-1	0.176	0.224	0.183	0.237	0.187	0.222	0.242	0.284
	TF∗IDF-2	0.047	0.154	0.049	0.182	0.047	0.167	0.097	0.233
	ROUGE-1	0.703	0.578	0.747	0.632	0.704	0.565	0.808	0.692
	ROUGE-2	0.695	0.572	0.718	0.635	0.727	0.583	0.803	0.694
	ROUGE-1-WE	0.571	0.450	0.579	0.458	0.586	0.437	0.653	0.516
	ROUGE-2-WE	0.566	0.397	0.556	0.388	0.607	0.413	0.671	0.481
	ROUGE-L	0.681	0.520	0.702	0.568	0.730	0.563	0.779	0.652
	FRAME-1	0.658	0.508	0.686	0.529	0.678	0.527	0.762	0.628
	FRAME-2	0.676	0.519	0.691	0.556	0.715	0.555	0.781	0.648
	BERTSCORE-F1	0.724	0.594	0.750	0.649	0.739	0.580	0.823	0.703
Word-Mover	WMD-1 + W2V	0.669	0.559	0.665	0.611	0.698	0.520	0.740	0.647
	WMD-1 + BERT + ROUTING	0.729	0.601	0.763	0.675	0.740	0.580	0.831	0.700
	WMD-1 + BERT + MNLI + ROUTING	0.734	0.609	0.768	0.686	0.747	0.589	0.837	0.711
	WMD-2 + BERT + MNLI + ROUTING	0.731	0.593	0.755	0.666	0.753	0.583	0.827	0.698
	WMD-1 + BERT + PMEANS	0.729	0.595	0.755	0.660	0.742	0.581	0.825	0.690
	WMD-1 + BERT + MNLI + PMEANS	0.736	0.604	0.760	0.672	0.754	0.594	0.831	0.701
	WMD-2 + BERT + MNLI + PMEANS	0.734	0.601	0.752	0.663	0.753	0.586	0.825	0.694

Table 8: Correlation of automatic metrics with summary-level human judgments for TAC-2008 and TAC-2009.

Setting	Metrics	BAGEL Inf	BAGEL Nat	BAGEL Qual	SFHOTEL Inf	SFHOTEL Nat	SFHOTEL Qual
Existing Metrics	BLEU-1	0.225	0.141	0.113	0.107	0.175	0.069
	BLEU-2	0.211	0.152	0.115	0.097	0.174	0.071
	BLEU-3	0.191	0.150	0.109	0.089	0.161	0.070
	BLEU-4	0.175	0.141	0.101	0.084	0.104	0.056
	ROUGE-L	0.202	0.134	0.111	0.092	0.147	0.062
	NIST	0.207	0.089	0.056	0.072	0.125	0.061
	CIDER	0.205	0.162	0.119	0.095	0.155	0.052
	METEOR	0.251	0.127	0.116	0.111	0.148	0.082
	BERTSCORE-F1	0.267	0.210	0.178	0.163	0.193	0.118
Word-WMD	WMD-1 + W2V	0.222	0.079	0.123	0.074	0.095	0.021
	WMD-1 + BERT + ROUTING	0.294	0.209	0.156	0.208	0.256	0.178
	WMD-1 + BERT + MNLI + ROUTING	0.278	0.180	0.144	0.211	0.252	0.175
	WMD-2 + BERT + MNLI + ROUTING	0.279	0.182	0.147	0.204	0.252	0.172
	WMD-1 + BERT + PMEANS	0.298	0.212	0.163	0.203	0.261	0.182
	WMD-1 + BERT + MNLI + PMEANS	0.285	0.195	0.158	0.207	0.270	0.183
	WMD-2 + BERT + MNLI + PMEANS	0.284	0.194	0.156	0.204	0.270	0.182

Table 9: Spearman correlation with utterance-level human judgments for BAGEL and SFHOTEL datasets.