MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance

Wei Zhao[†], **Maxime Peyrard**[†], **Fei Liu**[‡], **Yang Gao**[†], **Christian M. Meyer**[†], **Steffen Eger**[†]
[†] Computer Science Department, Technische Universität Darmstadt, Germany
[‡] Computer Science Department, University of Central Florida, US
zhao@aiphes.tu-darmstadt.de, maxime.peyrard@epfl.ch
feiliu@cs.ucf.edu, yang.gao@rhul.ac.uk
meyer@ukp.informatik.tu-darmstadt.de
eger@aiphes.tu-darmstadt.de

Abstract

A robust evaluation metric has a profound impact on the development of text generation systems. A desirable metric compares system output against references based on their semantics rather than surface forms. In this paper we investigate strategies to encode system and reference texts to devise a metric that shows a high correlation with human judgment of text quality. We validate our new metric, namely MoverScore, on a number of text generation tasks including summarization, machine translation, image captioning, and data-to-text generation, where the outputs are produced by a variety of neural and non-neural systems. Our findings suggest that metrics combining contextualized representations with a distance measure perform the best. Such metrics also demonstrate strong generalization capability across tasks. For ease-of-use we make our metrics available as web service.[1]

1 Introduction

The choice of evaluation metric has a significant impact on the assessed quality of natural language outputs generated by a system. A desirable metric assigns a single, real-valued score to the system output by comparing it with one or more reference texts for content matching. Many natural language generation (NLG) tasks can benefit from robust and unbiased evaluation, including text-to-text (*machine translation* and *summarization*), data-to-text (*response generation*), and image-to-text (*captioning*) (Gatt and Krahmer, 2018). Without proper evaluation, it can be difficult to judge on system competitiveness, hindering the development of advanced algorithms for text generation.

It is an increasingly pressing priority to develop better evaluation metrics given the recent advances in neural text generation. Neural models provide the flexibility to copy content from source text as well as generating unseen words (See et al., 2017). This aspect is hardly covered by existing metrics. With greater flexibility comes increased demand for unbiased evaluation. Diversity-promoting objectives make it possible to generate diverse natural language descriptions (Li et al., 2016; Wiseman et al., 2018). But standard evaluation metrics including BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) compute the scores based primarily on n-gram co-occurrence statistics, which are originally proposed for diagnostic evaluation of systems but not capable of evaluating text quality (Reiter, 2018), as they are not designed to measure if, and to what extent, the system and reference texts with distinct surface forms have conveyed the same meaning. Recent effort on the applicability of these metrics reveals that while compelling text generation system ascend on standard metrics, the text quality of system output is still hard to be improved (Böhm et al., 2019).

Our goal in this paper is to devise an automated evaluation metric assigning a single holistic score to any system-generated text by comparing it against human references for content matching. We posit that it is crucial to provide a holistic measure attaining high correlation with human judgments so that various neural and non-neural text generation systems can be compared directly. Intuitively, the metric assigns a perfect score to the system text if it conveys the same meaning as the reference text. Any deviation from the reference content can then lead to a reduced score, e.g., the system text contains more (or less) content than the reference, or the system produces ill-formed text that fails to deliver the intended meaning.

We investigate the effectiveness of a spectrum of distributional semantic representations to encode system and reference texts, allowing them to be compared for semantic similarity across

[1] Our code is publicly available at http://tiny.cc/vsqtbz

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 563–578, Hong Kong, China, November 3–7, 2019.

multiple natural language generation tasks. Our new metric quantifies the semantic distance between system and reference texts by harnessing the power of contextualized representations (Peters et al., 2018; Devlin et al., 2018) and a powerful distance metric (Rubner et al., 2000) for better content matching. Our contributions can be summarized as follows:

- We formulate the problem of evaluating generation systems as measuring the semantic distance between system and reference texts, assuming powerful continuous representations can encode any type of semantic and syntactic deviations.
- We investigate the effectiveness of existing contextualized representations and Earth Mover's Distance (Rubner et al., 2000) for comparing system predictions and reference texts, leading to our new automated evaluation metric that achieves high correlation with human judgments of text quality.
- Our metric outperforms or performs comparably to strong baselines on four text generation tasks including summarization, machine translation, image captioning, and data-to-text generation, suggesting this is a promising direction moving forward.

2 Related Work

It is of fundamental importance to design evaluation metrics that can be applied to natural language generation tasks of similar nature, including summarization, machine translation, data-to-text generation, image captioning, and many others. All these tasks involve generating texts of sentence or paragraph length. The system texts are then compared with one or more reference texts of similar length for semantic matching, whose scores indicate how well the systems perform on each task. In the past decades, however, evaluation of these natural language generation tasks has largely been carried out independently within each area.

Summarization A dominant metric for summarization evaluation is ROUGE (Lin, 2004), which measures the degree of lexical overlap between a system summary and a set of reference summaries. Its variants consider overlap of unigrams (-1), bigrams (-2), unigrams and skip bigrams with a maximum gap of 4 words (-SU4), longest common subsequences (-L) and its weighted version (-W-1.2), among others. Metrics such as Pyramid (Nenkova and Passonneau, 2004) and BE (Hovy et al., 2006; Tratz and Hovy, 2008) further compute matches of content units, e.g., (head-word, modifier) tuples, that often need to be manually extracted from reference summaries. These metrics achieve good correlations with human judgments in the past. However, they are not general enough to account for the relatedness between abstractive summaries and their references, as a system abstract can convey the same meaning using different surface forms. Furthermore, large-scale summarization datasets such as CNN/Daily Mail (Hermann et al., 2015) and Newsroom (Grusky et al., 2018) use a *single reference* summary, making it harder to obtain unbiased results when only lexical overlap is considered during summary evaluation.

Machine Translation A number of metrics are commonly used in MT evaluation. Most of these metrics compare system and reference translations based on surface forms such as word/character n-gram overlaps and edit distance, but not the meanings they convey. BLEU (Papineni et al., 2002) is a precision metric measuring how well a system translation overlaps with human reference translations using n-gram co-occurrence statistics. Other metrics include SentBLEU, NIST, chrF, TER, WER, PER, CDER, and METEOR (Lavie and Agarwal, 2007) that are used and described in the WMT metrics shared task (Bojar et al., 2017; Ma et al., 2018). RUSE (Shimanaka et al., 2018) is a recent effort to improve MT evaluation by training sentence embeddings on large-scale data obtained in other tasks. Additionally, preprocessing reference texts is crucial in MT evaluation, e.g., normalization, tokenization, compound splitting, etc. If not handled properly, different preprocessing strategies can lead to inconsistent results using word-based metrics (Post, 2018).

Data-to-text Generation BLEU can be poorly suited to evaluating data-to-text systems such as dialogue response generation and image captioning. These systems are designed to generate texts with lexical and syntactic variation, communicating the same information in many different ways. BLEU and similar metrics tend to reward systems that use the same wording as reference texts, causing repetitive word usage that is deemed undesirable to humans (Liu et al., 2016). In a similar vein, evaluating the quality of image captions can be challenging. CIDEr (Vedantam et al., 2015) uses tf-idf weighted n-grams for similarity estimation; and SPICE (Anderson et al., 2016) incorporates

synonym matching over scene graphs. Novikova et al. (2017) examine a large number of word- and grammar-based metrics and demonstrate that they only weakly reflect human judgments of system outputs generated by data-driven, end-to-end natural language generation systems.

Metrics based on Continuous Representations Moving beyond traditional metrics, we envision a new generation of automated evaluation metrics comparing system and reference texts based on semantics rather than surface forms to achieve better correlation with human judgments. A number of previous studies exploit static word embeddings (Ng and Abrecht, 2015; Lo, 2017) and trained classifers (Peyrard et al., 2017; Shimanaka et al., 2018) to improve semantic similarity estimation, replacing lexical overlaps.

In contemporaneous work, Zhang et al. (2019) describe a method comparing system and reference texts for semantic similarity leveraging the BERT representations (Devlin et al., 2018), which can be viewed as a special case of our metrics and will be discussed in more depth later. More recently, Clark et al. (2019) present a semantic metric relying on sentence mover's similarity and the ELMo representations (Peters et al., 2018) and apply them to summarization and essay scoring. Mathur et al. (2019) introduce unsupervised and supervised metrics based on the BERT representations to improve MT evaluation, while Peyrard (2019a) provides a composite score combining redundancy, relevance and informativeness to improve summary evaluation.

In this paper, we seek to accurately measure the (dis)similarity between system and reference texts drawing inspiration from contextualized representations and Word Mover's Distance (WMD; Kusner et al., 2015). WMD finds the "traveling distance" of moving from the word frequency distribution of the system text to that of the reference, which is essential to capture the (dis)similarity between two texts. Our metrics differ from the contemporaneous work in several facets: (i) we explore the granularity of embeddings, leading to two variants of our metric, word mover and sentence mover; (ii) we investigate the effectiveness of diverse pretrained embeddings and finetuning tasks; (iii) we study the approach to consolidate layer-wise information within contextualized embeddings; (iii) our metrics demonstrate strong generalization capability across four tasks, oftentimes outperforming the supervised ones. We now describe our method in detail.

3 Our MoverScore Meric

We have motivated the need for better metrics capable of evaluating disparate NLG tasks. We now describe our metric, namely MoverScore, built upon a combination of (i) contextualized representations of system and reference texts and (ii) a distance between these representations measuring the semantic distance between system outputs and references. It is particularly important for a metric to not only capture the amount of *shared content* between two texts, i.e., intersect(A,B), as is the case with many semantic textual similarity measures (Peters et al., 2018; Devlin et al., 2018); but also to accurately reflect to what extent the system text has *deviated* from the reference, i.e., union(A,B) - intersect(A,B), which is the intuition behind using a distance metric.

3.1 Measuring Semantic Distance

Let $\boldsymbol{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_m)$ be a sentence viewed as a sequence of words. We denote by $\boldsymbol{x}^n$ the sequence of n-grams of $\boldsymbol{x}$ (i.e., $\boldsymbol{x}^1 = \boldsymbol{x}$ is the sequence of words and $\boldsymbol{x}^2$ is the sequence of bigrams). Furthermore, let $\boldsymbol{f}_{\boldsymbol{x}^n} \in \mathbb{R}_+^{|\boldsymbol{x}^n|}$ be a vector of weights, one weight for each n-gram of $\boldsymbol{x}^n$. We can assume $\boldsymbol{f}_{\boldsymbol{x}^n}^\intercal \mathbf{1} = 1$, making $\boldsymbol{f}_{\boldsymbol{x}^n}$ a distribution over n-grams. Intuitively, the effect of some n-grams like those including function words can be downplayed by giving them lower weights, e.g., using Inverse Document Frequency (IDF).

Word Mover's Distance (WMD) (Kusner et al., 2015), a special case of Earth Mover's Distance (Rubner et al., 2000), measures semantic distance between texts by aligning semantically similar words and finding the amount of flow traveling between these words. It was shown useful for text classification and textual similarity tasks (Kusner et al., 2015). Here, we formulate a generalization operating on n-grams. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two sentences viewed as sequences of n-grams: $\boldsymbol{x}^n$ and $\boldsymbol{y}^n$. If we have a distance metric d between n-grams, then we can define the *transportation cost* matrix $\boldsymbol{C}$ such that $\boldsymbol{C}_{ij} = d(\mathrm{x}_i^n, \mathrm{y}_j^n)$ is the distance between the i-th n-gram of $\boldsymbol{x}$ and the j-th n-gram of $\boldsymbol{y}$. The WMD between the two sequences of n-grams $\boldsymbol{x}^n$ and $\boldsymbol{y}^n$ with associated n-gram weights $\boldsymbol{f}_{\boldsymbol{x}^n}$ and $\boldsymbol{f}_{\boldsymbol{y}^n}$ is then given by:

$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$
$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^\intercal \mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

where $\boldsymbol{F}$ is the *transportation flow* matrix with $\boldsymbol{F}_{ij}$ denoting the amount of flow traveling from the i-th n-gram x_i^n in $\boldsymbol{x}^n$ to the j-th n-gram y_j^n in $\boldsymbol{y}^n$. Here, $\langle \boldsymbol{C}, \boldsymbol{F} \rangle$ denotes the sum of all matrix entries of the matrix $\boldsymbol{C} \odot \boldsymbol{F}$, where $\odot$ denotes element-wise multiplication. Then $\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n)$ is the minimal transportation cost between $\boldsymbol{x}^n$ and $\boldsymbol{y}^n$ where n-grams are weighted by $\boldsymbol{f}_{\boldsymbol{x}^n}$ and $\boldsymbol{f}_{\boldsymbol{y}^n}$.

In practice, we compute the Euclidean distance between the embedding representations of n-grams: $d(\mathrm{x}_i^n, \mathrm{y}_j^n) = ||E(\mathrm{x}_i^n) - E(\mathrm{y}_j^n)||_2$ where E is the embedding function which maps an n-gram to its vector representation. Usually, *static* word embeddings like word2vec are used to compute E but these cannot capture word order or compositionality. Alternatively, we investigate contextualized embeddings like ELMo and BERT because they encode information about the whole sentence into each word vector.

We compute the n-gram embeddings as the weighted sum over its word embeddings. Formally, if $\boldsymbol{x}_i^n = (\mathrm{x}_i, \ldots, \mathrm{x}_{i+n-1})$ is the i-th n-gram from sentence $\boldsymbol{x}$, its embedding is given by:

$$E(\boldsymbol{x}_i^n) = \sum_{k=i}^{i+n-1} \mathrm{idf}(\mathrm{x}_k) \cdot E(\mathrm{x}_k) \qquad (1)$$

where $\mathrm{idf}(\mathrm{x}_k)$ is the IDF of word x_k computed from all sentences in the corpus and $E(\mathrm{x}_k)$ is its word vector. Furthermore, the weight associated to the n-gram $\boldsymbol{x}_i^n$ is given by:

$$\boldsymbol{f}_{\boldsymbol{x}_i^n} = \frac{1}{Z} \sum_{k=i}^{i+n-1} \mathrm{idf}(\mathrm{x}_k) \qquad (2)$$

where Z is a normalizing constant s.t. $\boldsymbol{f}_{\boldsymbol{x}^n}^\top \mathbf{1} = 1$,

In the limiting case where n is larger than the sentence's size, $\boldsymbol{x}^n$ contains only one n-gram: the whole sentence. Then $\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n)$ reduces to computing the distance between the two sentence embeddings, namely Sentence Mover's Distance (SMD), denoted as:

$$\mathrm{SMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := ||E(\boldsymbol{x}_1^{l_x}) - E(\boldsymbol{y}_1^{l_y})||$$

where l_x and l_y are the size of sentences.

Hard and Soft Alignments In contemporaneous work, BERTScore (Zhang et al., 2019) also models the semantic distance between system and reference texts for evaluating text generation systems. As shown in Figure 1, BERTScore (precision/recall) can be intuitively viewed as hard alignments (one-to-one) for words in a sentence pair, where each word in one sequence travels to the most semantically similar word in the other sequence. In contrast, MoverScore goes beyond BERTScore as it relies on soft alignments (many-to-one) and allows to map semantically related words in one sequence to the respective word in the other sequence by solving a constrained optimization problem: finding the minimum effort to transform between two texts.

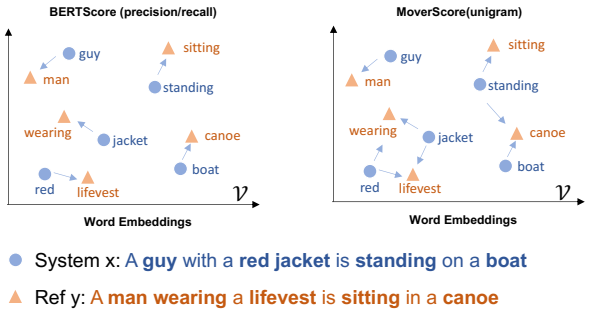

Figure 1: An illustration of MoverScore and BERTScore.

The formulation of Word Mover's Distance provides an important possibility to bias the metric towards precision or recall by using an asymmetric transportation cost matrix, which bridges a gap between MoverScore and BERTScore:

Proposition 1 *BERTScore (precision/recall) can be represented as a (non-optimized) Mover Distance $\langle \boldsymbol{C}, \boldsymbol{F} \rangle$, where $\boldsymbol{C}$ is a transportation cost matrix based on BERT and $\boldsymbol{F}$ is a uniform transportation flow matrix.*[2]

3.2 Contextualized Representations

The task formulation naturally lends itself to deep contextualized representations for inducing word vectors $E(\mathrm{x}_i)$. Despite the recent success of multi-layer attentive neural architectures (Devlin et al., 2018; Peters et al., 2018), consolidating layer-wise information remains an open problem as different layers capture information at disparate scales and task-specific layer selection methods may be limited (Liu et al., 2018, 2019). Tenney et al. (2019) found that a scalar mix of output layers trained from task-dependent supervisions would be effective in a deep transformer-based model. Instead, we investigate aggregation functions to consolidate layer-wise information, forming stationary representations of words without supervision.

Consider a sentence $\boldsymbol{x}$ passed through contextualized encoders such as ELMo and BERT with L layers. Each layer of the encoders produces a vec-

[2] See the proof in the appendix.

tor representation for each word x_i in $\boldsymbol{x}$. We denote by $\boldsymbol{z}_{i,l} \in \mathbb{R}^d$ the representation given by layer l, a d-dimensional vector. Overall, x_i receives L different vectors $(\boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,L})$. An aggregation ϕ maps these L vectors to one final vector:

$$E(x_i) = \phi(\boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,L}) \quad (3)$$

where $E(x_i)$ is the aggregated representation of the word x_i.

We study two alternatives for ϕ: (i) the concatenation of power means (Rücklé et al., 2018) as a generalized pooling mechanism, and (ii) a routing mechanism for aggregation (Zhao et al., 2018, 2019). We relegate the routing method to appendix, as it does not yield better results than power means.

Power Means Power means is an effective generalization of pooling techniques for aggregating information. It computes a non-linear average of a set of values with an exponent p (Eq. (4)). Following Rücklé et al. (2018), we exploit power means to aggregate vector representations $(\boldsymbol{z}_{i,l})_{l=1}^{L}$ pertaining to the i-th word from all layers of a deep neural architecture. Let $p \in \mathbb{R} \cup \{\pm\infty\}$, the p-mean of $(\boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,L})$ is:

$$\boldsymbol{h}_i^{(p)} = \left(\frac{\boldsymbol{z}_{i,1}^p + \cdots + \boldsymbol{z}_{i,L}^p}{L} \right)^{1/p} \in \mathbb{R}^d \quad (4)$$

where exponentiation is applied elementwise. This generalized form can induce common named means such as arithmetic mean ($p = 1$) and geometric mean ($p = 0$). In extreme cases, a power mean reduces to the minimum value of the set when $p = -\infty$, and the maximum value when $p = +\infty$. The concatenation of p-mean vectors we use in this paper is denoted by:

$$E(x_i) = \mathbf{h}_i^{(p_1)} \oplus \cdots \oplus \mathbf{h}_i^{(p_K)} \quad (5)$$

where $\oplus$ is vector concatenation; $\{p_1, \ldots, p_K\}$ are exponent values, and we use $K = 3$ with $p = 1, \pm\infty$ in this work.

3.3 Summary of MoverScore Variations

We investigate our MoverScore along four dimensions: (i) the granularity of embeddings, i.e., the size of n for n-grams, (ii) the choice of pretrained embedding mechanism, (iii) the fine-tuning task used for BERT[3] (iv) the aggregation technique (p-means or routing) when applicable.

[3]ELMo usually requires heavy layers on the top, which restricts the power of fine-tuning tasks for ELMo.

Granularity We used $n = 1$ and $n = 2$ as well as full sentences ($n =$ size of the sentence).

Embedding Mechanism We obtained word embeddings from three different methods: *static* embedding with word2vec as well as contextualized embedding with ELMo and BERT. If $n > 1$, n-gram embeddings are calculated by Eq. (1). Note that they represent sentence embeddings when $n =$ size of the sentence.

Fine-tuning Tasks Natural Language Inference (NLI) and paraphrasing pose high demands in understanding sentence meaning. This motivated us to fine-tune BERT representations on two NLI datasets, MultiNLI and QANLI, and one Paraphrase dataset, QQP—the largest datasets in GLUE (Wang et al., 2018). We fine-tune BERT on each of these, yielding different contextualized embeddings for our general evaluation metrics.

Aggregation For ELMo, we aggregate word representations given by all three ELMo layers, using p-means or routing (see the appendix). Word representations in BERT are aggregated from the last five layers, using p-means or routing since the representations in the initial layers are less suited for use in downstream tasks (Liu et al., 2019).

4 Empirical Evaluation

In this section, we measure the quality of different metrics on four tasks: *machine translation*, *text summarization*, *image captioning* and *dialogue generation*. Our major focus is to study the correlation between different metrics and human judgment. We employ two text encoders to embed n-grams: $\text{BERT}_{\text{base}}$, which uses a 12-layer transformer, and $\text{ELMO}_{\text{original}}$, which uses a 3-layer BiLSTM. We use Pearson's r and Spearman's ρ to measure the correlation. We consider two variants of MoverScore: *word mover* and *sentence mover*, described below.

Word Mover We denote our word mover notation containing four ingredients as: *WMD-Granularity+Embedding+Finetune+Aggregation*. For example, WMD-1+BERT+MNLI+PMEANS represents the semantic metric using word mover distance where unigram-based word embeddings fine-tuned on MNLI are aggregated by p-means.

Sentence Mover We denote our sentence mover notation with three ingredients as: *SMD+Embedding+Finetune+Aggregation*. For example, SMD+W2V represents the semantic

metric using sentence mover distance where two sentence embeddings are computed as the weighted sum over their word2vec embeddings by Eq. (1).

Baselines We select multiple strong baselines for each task for comparison: SentBLEU, METEOR++ (Guo et al., 2018), and a supervised metric RUSE for machine translation; ROUGE-1 and ROUGE-2 and a supervised metric S^3_{best} (Peyrard et al., 2017) for text summarization; BLEU and METEOR for dialogue response generation, CIDEr, SPICE, METEOR and a supervised metric LEIC (Cui et al., 2018) for image captioning. We also report BERTScore (Zhang et al., 2019) for all tasks (see §2). Due to the page limit, we only compare with the strongest baselines, the rest can be found in the appendix.

4.1 Machine Translation

Data We obtain the source language sentences, their system and reference translations from the WMT 2017 news translation shared task (Bojar et al., 2017). We consider 7 language pairs: from German (de), Chinese (zh), Czech (cs), Latvian (lv), Finnish (fi), Russian (ru), and Turkish (tr), resp. to English. Each language pair has approximately 3,000 sentences, and each sentence has one reference translation and multiple system translations generated by participating systems. For each system translation, at least 15 human assessments are independently rated for quality.

Results Table 1: In all language pairs, the best correlation is achieved by our word mover metrics that use a BERT pretrained on MNLI as the embedding generator and PMeans to aggregate the embeddings from different BERT layers, i.e., WMD-1/2+BERT+MNLI+PMeans. Note that our unsupervised word mover metrics even outperforms RUSE, a supervised metric. We also find that our word mover metrics outperforms the sentence mover. We conjecture that important information is lost in such a sentence representation while transforming the whole sequence of word vectors into one sentence embedding by Eq. (1).

4.2 Text Summarization

We use two summarization datasets from the Text Analysis Conference (TAC)[4]: TAC-2008 and TAC-2009, which contain 48 and 44 *clusters*, respectively. Each cluster includes 10 news articles (on the same topic), four reference summaries, and 57 (in TAC-2008) or 55 (in TAC-2009) system summaries generated by the participating systems. Each summary (either reference or system) has fewer than 100 words, and receives two human judgment scores: the *Pyramid* score (Nenkova and Passonneau, 2004) and the *Responsiveness* score. Pyramid measures how many important semantic content units in the reference summaries are covered by the system summary, while Responsiveness measures how well a summary responds to the overall quality combining both content and linguistic quality.

Results Tables 2: We observe that lexical metrics like ROUGE correlate above-moderate on TAC 2008 and 2009 datasets. In contrast, these metrics perform poorly on other tasks like Dialogue Generation (Novikova et al., 2017) and Image Captioning (Anderson et al., 2016). Apparently, strict matches on surface forms seems reasonable for *extractive* summarization datasets. However, we still see that our word mover metrics, i.e., WMD-1+BERT+MNLI+PMeans, perform better than or come close to even the supervised metric S^3_{best}.

4.3 Data-to-text Generation

We use two task-oriented dialogue datasets: BAGEL (Mairesse et al., 2010) and SFHOTEL (Wen et al., 2015), which contains 202 and 398 instances of Meaning Representation (MR). Each MR instance includes multiple references, and roughly two system utterances generated by different neural systems. Each system utterance receives three human judgment scores: *informativeness*, *naturalness* and *quality* score (Novikova et al., 2017). Informativeness measures how much information a system utterance provides with respect to an MR. Naturalness measures how likely a system utterance is generated by native speakers. Quality measures how well a system utterance captures fluency and grammar.

Results Tables 3: Interestingly, no metric produces an even moderate correlation with human judgments, including our own. We speculate that current contextualizers are poor at representing named entities like hotels and place names as well as numbers appearing in system and reference texts. However, best correlation is still achieved by our word mover metrics combining contextualized representations.

[4] http://tac.nist.gov

Setting	Metrics	Direct Assessment cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
BASELINES	METEOR++	0.552	0.538	0.720	0.563	0.627	0.626	0.646	0.610
	RUSE(*)	0.624	0.644	0.750	0.697	0.673	0.716	0.691	0.685
	BERTSCORE-F1	0.670	0.686	0.820	0.710	0.729	0.714	0.704	0.719
SENT-MOVER	SMD + W2V	0.438	0.505	0.540	0.442	0.514	0.456	0.494	0.484
	SMD + ELMO + PMEANS	0.569	0.558	0.732	0.525	0.581	0.620	0.584	0.595
	SMD + BERT + PMEANS	0.607	0.623	0.770	0.639	0.667	0.641	0.619	0.652
	SMD + BERT + MNLI + PMEANS	0.616	0.643	0.785	0.660	0.664	0.668	0.633	0.667
WORD-MOVER	WMD-1 + W2V	0.392	0.463	0.558	0.463	0.456	0.485	0.481	0.471
	WMD-1 + ELMO + PMEANS	0.579	0.588	0.753	0.559	0.617	0.679	0.645	0.631
	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 1: Absolute Pearson correlations with segment-level human judgments in 7 language pairs on WMT17 dataset.

Setting	Metrics	TAC-2008 Responsiveness r	Responsiveness ρ	Pyramid r	Pyramid ρ	TAC-2009 Responsiveness r	Responsiveness ρ	Pyramid r	Pyramid ρ
BASELINES	S^3_{best} (*)	0.715	0.595	0.754	0.652	0.738	**0.595**	**0.842**	**0.731**
	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
	BERTSCORE-F1	0.724	0.594	0.750	0.649	0.739	0.580	0.823	0.703
SENT-MOVER	SMD + W2V	0.583	0.469	0.603	0.488	0.577	0.465	0.670	0.560
	SMD + ELMO + PMEANS	0.631	0.472	0.631	0.499	0.663	0.498	0.726	0.568
	SMD + BERT + PMEANS	0.658	0.530	0.664	0.550	0.670	0.518	0.731	0.580
	SMD + BERT + MNLI + PMEANS	0.662	0.525	0.666	0.552	0.667	0.506	0.723	0.563
WORD-MOVER	WMD-1 + W2V	0.669	0.549	0.665	0.588	0.698	0.520	0.740	0.647
	WMD-1 + ELMO + PMEANS	0.707	0.554	0.726	0.601	0.736	0.553	0.813	0.672
	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 2: Pearson r and Spearman ρ correlations with summary-level human judgments on TAC 2008 and 2009.

4.4 Image Captioning

We use a popular image captioning dataset: MS-COCO (Lin et al., 2014), which contains 5,000 images. Each image includes roughly five reference captions, and 12 system captions generated by the participating systems from 2015 COCO Captioning Challenge. For the system-level human correlation, each system receives five human judgment scores: M1, M2, M3, M4, M5 (Anderson et al., 2016). The M1 and M2 scores measure overall quality of the captions while M3, M4 and M5 scores measure correctness, detailedness and saliency of the captions. Following Cui et al. (2018), we compare the Pearson correlation with two system-level scores: M1 and M2, since we focus on studying metrics for the overall quality of the captions, leaving metrics understanding captions in different aspects (correctness, detailedness and saliency) to future work.

Results Table 4: Word mover metrics outperform all baselines except for the supervised metric LEIC, which uses more information by considering *both* images and texts.

4.5 Further Analysis

Hard and Soft Alignments BERTScore is the harmonic mean of BERTScore-Precision and BERTScore-Recall, where both two can be decomposed as a combination of "Hard Mover Distance" (HMD) and BERT (see Prop. 1).

We use the representations in the 9-th BERT layer for fair comparison of BERTScore and MoverScore and show results on the machine translation task in Table 5. MoverScore outperforms both asymmetric HMD factors, while if they are combined via harmonic mean, BERTScore is on par with MoverScore. We conjecture that BERT softens hard alignments of BERTScore as contextualized embeddings encode information about the whole sentence into each word vector. We also observe that WMD-BIGRAMS slightly outperforms WMD-UNIGRAMS on 3 out of 4 language pairs.

Setting	Metrics	BAGEL Inf	BAGEL Nat	BAGEL Qual	SFHOTEL Inf	SFHOTEL Nat	SFHOTEL Qual
BASELINES	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
	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
SENT-MOVER	SMD + W2V	0.024	0.074	0.078	0.022	0.025	0.011
	SMD + ELMO + PMEANS	0.251	0.171	0.147	0.130	0.176	0.096
	SMD + BERT + PMEANS	0.290	0.163	0.121	0.192	0.223	0.134
	SMD + BERT + MNLI + PMEANS	0.280	0.149	0.120	0.205	0.239	0.147
WORD-MOVER	WMD-1 + W2V	0.222	0.079	0.123	0.074	0.095	0.021
	WMD-1 + ELMO + PMEANS	0.261	0.163	0.148	0.147	0.215	0.136
	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 3: Spearman correlation with utterance-level human judgments for BAGEL and SFHOTEL datasets.

Setting	Metric	M1	M2
BASELINES	LEIC(*)	**0.939**	**0.949**
	METEOR	0.606	0.594
	SPICE	0.759	0.750
	BERTSCORE-RECALL	0.809	0.749
SENT-MOVER	SMD + W2V	0.683	0.668
	SMD + ELMO + P	0.709	0.712
	SMD + BERT + P	0.723	0.747
	SMD + BERT + M + P	0.789	0.784
WORD-MOVER	WMD-1 + W2V	0.728	0.764
	WMD-1 + ELMO + P	0.753	0.775
	WMD-1 + BERT + P	0.780	0.790
	WMD-1 + BERT + M + P	**0.813**	**0.810**
	WMD-2 + BERT + M + P	0.812	0.808

Table 4: Pearson correlation with system-level human judgments on MSCOCO dataset. 'M' and 'P' are short names.

Metrics	cs-en	de-en	fi-en	lv-en
RUSE	0.624	0.644	0.750	0.697
HMD-F1 + BERT	0.655	0.681	0.821	0.712
HMD-RECALL + BERT	0.651	0.658	0.788	0.681
HMD-PREC + BERT	0.624	0.669	0.817	0.707
WMD-UNIGRAM + BERT	0.651	0.686	**0.823**	0.710
WMD-BIGRAM + BERT	**0.665**	**0.688**	0.821	**0.712**

Table 5: Comparison on hard and soft alignments.

Distribution of Scores In Figure 2, we take a closer look at sentence-level correlation in MT. Results reveal that the lexical metric SENTBLEU can correctly assign lower scores to system translations of low quality, while it struggles in judging system translations of high quality by assigning them lower scores. Our finding agrees with the observations found in Chaganty et al. (2018); Novikova et al. (2017): lexical metrics correlate better with human judgments on texts of low quality than high quality. Peyrard (2019b) further show that lexical metrics cannot be trusted because

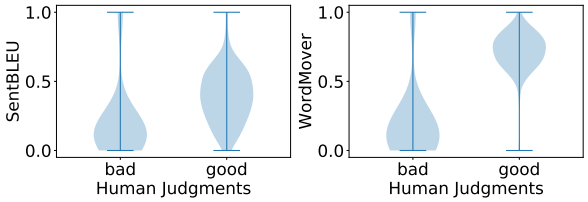

Figure 2: Score distribution in German-to-English pair.

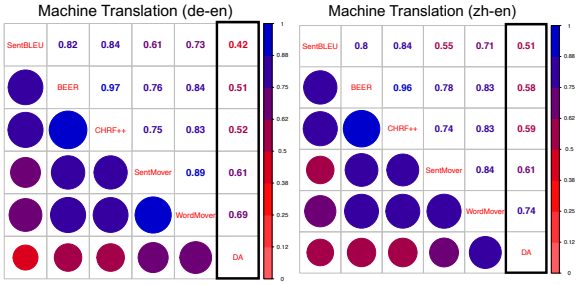

Figure 3: Correlation in similar language (de-en) and distant language (zh-en) pair, where bordered area shows correlations between human assessment and metrics, the rest shows inter-correlations across metrics and DA is direct assessment rated by language experts.

they strongly disagree on high-scoring system outputs. Importantly, we observe that our word mover metric combining BERT can clearly distinguish texts of two polar qualities.

Correlation Analysis In Figure 3, we observe existing metrics for MT evaluation attaining medium correlations (0.4-0.5) with human judgments but high inter-correlations between themselves. In contrast, our metrics can attain high correlations (0.6-0.7) with human judgments, performing robust across different language pairs. We believe that our improvements come from clearly distinguishing translations that fall on two extremes.

Impact of Fine-tuning Tasks Figure 4 com-

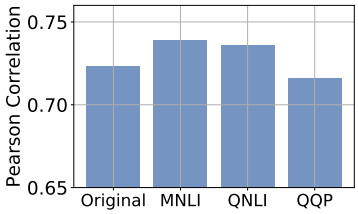

Figure 4: Correlation is averaged over 7 language pairs.

pares Pearson correlations with our word mover metrics combining BERT fine-tuned on three different tasks. We observe that fine-tuning on closely related tasks improves correlations, especially fine-tuning on MNLI leads to an impressive improvement by 1.8 points on average.

4.6 Discussions

We showed that our metric combining contextualized embeddings and Earth Mover's Distance outperforms strong unsupervised metrics on 3 out of 4 tasks, i.e., METEOR++ on machine translation by 5.7 points, SPICE on image captioning by 3.0 points, and METEOR on dialogue response generation by 2.2 points. The best correlation we achieved is combining contextualized word embeddings and WMD, which even rivals or exceeds SOTA task-dependent *supervised* metrics across different tasks. Especially in machine translation, our word mover metric pushes correlations in machine translation to 74.3 on average (5.8 points over the SOTA supervised metric and 2.4 points over contemporaneous BERTScore). The major improvements come from contextualized BERT embeddings rather than word2vec and ELMo, and from fine-tuning BERT on large NLI datasets. However, we also observed that soft alignments (MoverScore) marginally outperforms hard alignments (BERTScore). Regarding the effect of n-grams in word mover metrics, unigrams slightly outperforms bigrams on average. For the effect of aggregation functions, we suggested effective techniques for layer-wise consolidations, namely p-means and routing, both of which are close to the performance of the best layer and on par with each other (see the appendix).

5 Conclusion

We investigated new unsupervised evaluation metrics for text generation systems combining contextualized embeddings with Earth Mover's Distance. We experimented with two variants of our metric, sentence mover and word mover. The latter has demonstrated strong generalization ability across four text generation tasks, oftentimes even outperforming supervised metrics. Our metric provides a promising direction towards a holistic metric for text generation and a direction towards more 'human-like' (Eger et al., 2019) evaluation of text generation systems.

In future work, we plan to avoid the need for costly human references in the evaluation of text generation systems, and instead base evaluation scores on source texts and system predictions only, which would allow for 'next-level', unsupervised (in a double sense) and unlimited evaluation (Louis and Nenkova, 2013; Böhm et al., 2019).

Acknowledgments

We thank the anonymous reviewers for their comments, which greatly improved the final version of the paper. This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1. Fei Liu is supported in part by NSF grant IIS-1909603.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 382–398.

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*, Hong Kong, China.

Ondrej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Conference on Machine Translation (WMT)*.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653.

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.

Dorin Comaniciu and Peter Meer. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):603–619.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5804–5812.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. 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 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research (JAIR)*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. NEWSROOM: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Yinuo Guo, Chong Ruan, and Junfeng Hu. 2018. Meteor++: Incorporating copy knowledge into machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 740–745.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of Neural Information Processing Systems (NIPS)*.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006)*, pages 604–611.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of ACL workshop on Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Liyuan Liu, Xiang Ren, Jingbo Shang, Xiaotao Gu, Jian Peng, and Jiawei Han. 2018. Efficient contextualized representation: Language model pruning for sequence labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 589–597.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Qingsong Ma, Ondrej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation

using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561. Association for Computational Linguistics.

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.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 145–152. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Maxime Peyrard. 2019a. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard. 2019b. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

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*.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *CoRR*, abs/1803.01400.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Stephen Tratz and Eduard H Hovy. 2008. Summarization Evaluation Using Transformed Basic Elements. In *Proceedings of the text analysing conference, (TAC 2008)*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575.

Matt P Wand and M Chris Jones. 1994. *Kernel smoothing*. Chapman and Hall/CRC.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2018. Learning neural templates for text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Suofei Zhang, Wei Zhao, Xiaofu Wu, and Quan Zhou. 2018. Fast dynamic routing based on weighted kernel density estimation. *CoRR*, abs/1805.10807.

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

Wei Zhao, Haiyun Peng, Steffen Eger, Erik Cambria, and Min Yang. 2019. Towards scalable and reliable capsule networks for challenging NLP applications. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1549–1559, Florence, Italy. Association for Computational Linguistics.

Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Suofei Zhang, and Zhou Zhao. 2018. Investigating capsule networks with dynamic routing for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

A Supplemental Material

A.1 Proof of Prop. 1

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

As a reminder, the WMD formulation is:

$$\text{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \sum_{i,j} \boldsymbol{C}_{ij} \cdot \boldsymbol{F}_{ij}$$
$$\text{s.t. } \mathbf{1}^\intercal \boldsymbol{F}^\intercal \mathbf{1} = 1, \quad \mathbf{1}^\intercal \boldsymbol{F} \mathbf{1} = 1.$$

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

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:

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

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 (omitted). Then, F_{BERT} can be formulated as harmonic-mean of two WMD forms of P_{BERT} and R_{BERT}.

A.2 Routing

In this section, we study the aggregation function ϕ with a routing scheme, which has achieved good results in other NLP tasks (Zhao et al., 2018, 2019). 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:

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

where $d(\cdot)$ is a distance function, γ_{ij} denotes the underlying closeness between the aggregated vector $\boldsymbol{v}_j$ and vector $\boldsymbol{z}_i$ in the i-th layer, and k is a kernel function. Some instantiations of $k(\cdot)$ (Wand and Jones, 1994) are:

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

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 $\gamma_{ij}^{\tau+1}$ 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 a hyperparameter to control step size. The routing process is summarized in Algorithm 1.

Algorithm 1 Aggregation by Routing

procedure ROUTING($\boldsymbol{z}_{ij}, \ell$)
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 (using p-means and routing). We often see that which layer has best performance is task-dependent, and our word mover based metrics (WMD) with p-means or routing schema come close to the oracle performance obtained from the best layers.

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, respectively. We find that word mover based metrics combining BERT fine-tuned on MNLI have highest correlations with humans, outperforming all of the unsupervised metrics and even supervised metrics like RUSE and S^3_{full}. Routing and p-means perform roughly equally well.

Metrics	Direct Assessment 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.

Setting	Metrics	Direct Assessment cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
BASELINES	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
BASELINES	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			SFHOTEL		
		Inf	**Nat**	**Qual**	**Inf**	**Nat**	**Qual**
BASELINES	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-MOVER	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.