Automatic Text Evaluation through the Lens of Wasserstein Barycenters

Pierre Colombo[1,2], Guillaume Staerman[1], Chloe Clavel[1], Pablo Piantanida[*]

[1]LTCI, Telecom Paris, Institut Polytechnique de Paris,
[2]IBM GBS France,
[*]CentraleSupelec-CNRS, Universite Paris-Saclay
[1]firstname.lastname@telecom-paris.fr,
[2]pierre.colombo@ibm.com,
[*]pablo.piantanida@centralesupelec.fr,

Abstract

A new metric `BaryScore` to evaluate text generation based on deep contextualized embeddings (*e.g.*, BERT, Roberta, ELMo) is introduced. This metric is motivated by a new framework relying on optimal transport tools, *i.e.*, Wasserstein distance and barycenter. By modelling the layer output of deep contextualized embeddings as a probability distribution rather than by a vector embedding; this framework provides a natural way to aggregate the different outputs through the Wasserstein space topology. In addition, it provides theoretical grounds to our metric and offers an alternative to available solutions (*e.g.*, MoverScore and BertScore). Numerical evaluation is performed on four different tasks: machine translation, summarization, data2text generation and image captioning. Our results show that `BaryScore` outperforms other BERT based metrics and exhibits more consistent behaviour in particular for text summarization.

1 Introduction

Automatic Evaluation (AE) of Natural Language Generation (NLG) is a key problem towards better systems (Specia et al., 2010). It allows to assess the quality of generated text without relying on human evaluation campaigns that are expensive and time consuming (Belz and Reiter, 2006; Sai et al., 2020). For instance, it becomes crucial to design automatic and effective metrics with simultaneous goals: (i) to be able to compare, to control and to debug systems without relying on human annotators (Peyrard, 2019a,b); and (ii) to improve the learning phase of models by deriving losses that are better surrogate of human judgment than the widely used cross-entropy loss (Clark et al., 2019).

A plethora of automatic metrics has been introduced these last few years and may be grouped into two general classes: trained (Ma et al., 2017; Shimanaka et al., 2018; Lowe et al., 2016; Lita et al., 2005) and untrained metrics (Doddington, 2002;

Popović, 2015). In this paper, we mainly focus on untrained metrics that can be further split into three subgroups: *string matching* (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005; Doddington, 2002; Popović, 2015), *edit based* (Leusch et al., 2006; Snover et al., 2006; Wang et al., 2016) and *embedding based* metrics (Chow et al., 2019; Kusner et al., 2015; Lo and Wu, 2011; Lo, 2019). Both string matching and edit based metrics fail to assign reliable scores when reference and candidate convey the same meaning with distinct surface forms (Reiter and Belz, 2009) (*e.g.*, case of synonyms and paraphrases). These shortcomings have been addressed by metrics based on continuous representations. Recently, they have benefited from contextual embeddings such as ELMO (Peters et al., 2018) and BERT (Devlin et al., 2018; Liu et al., 2019b). Perhaps, the most known are BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019b) or Sentence Mover (Clark et al., 2019) which optimise Word Mover Distance (WMD) (Kusner et al., 2015)[1], a particular instance of optimal transport (OT) problem.

Originally introduced by Kusner et al. (2015) the WMD is used to compute the Wasserstein distance between text documents relying on a *single layer* embedding such as GloVe (Pennington et al., 2014) or Wor2Vect (Mikolov et al., 2013). To apply WMD with *multi layers* embedding several recipes have been proposed. BertScore selects the best layer based on a validation set. However, the selection of the validation set is arbitrary while on the other hand, a single layer selection does not exploit the information available in other layers (Voita et al., 2019; Hewitt and Liang, 2019; Liu et al., 2019a). MoverScore and Sentence Mover attempt to leverage the information available in other layers by aggregating the layers using a power mean (Rücklé et al., 2018b). In addition to adding extra

---

[1](Zhao et al., 2019b) shows that BertScore can be viewed as a non-optimized transport problem.

hyper-parameters, this aggregation method relies on euclidean topology which induces a geometrical discrepancy as the final cost is computed using a Wasserstein distance.

Our contributions. We introduce `BaryScore` a novel metric, which addresses aforementioned aggregation pitfalls by relying on Wasserstein barycenters and evaluate its performance on four different tasks: neural machine translation, text summarization, image captioning and data2text generation. Our main contributions can be summarized as follow:

1. *A novel metric to measure the semantic equivalence between two texts.* This metric relies on the embedding geometry of the layers induced in Wasserstein spaces. In order to overcome the geometric distortion generated by the aggregation techniques used to compute WMD with deep embeddings (*e.g.*, BERT, ELMo and Roberta), we aggregate layers information using the Wasserstein barycenter. This new formulation offers a topological advantage, *i.e.*, using barycenters giving meaning to the use of OT based distance afterwards and is parameter-free, i.e., it avoids choosing by hand the best layer (as for BertScore) or selecting the exponent in the power means (as for MoverScore). Our formulation provides an alternative and a generalization to the WMD formulation (Kusner et al., 2015) (originally introduced for Word2Vec) when applied to embeddings which are coming from multi-layer neural networks and thus, it provides theoretical motivation to `BaryScore`[2] a new metric that aggregates deep contextualized embedding using Wasserstein barycenters.

2. *Applications and numerical results.* We demonstrate that `BaryScore` provides better results than a large variety of state-of-art untrained metrics on four text generation tasks: namely NMT, summarization, image captioning, data2text suggesting that Wasserstein barycenters offer a promising direction moving forward.

2   Related Work

The goal of NLG is to generate coherent, readable and informative text from some input data (*e.g.*, texts, images and tables). However, the exact definition of each of these three criteria remains task-dependent and thus, making it hard to provide a unique metric for all tasks. As an example, NMT focuses on fluency, fidelity and adaquatie (Hovy,

---

²Bary stands for Barycenter.

1999; White et al., 1994) in contrast to summarization where annotators have to focus on coherence, content, readability, grammatically, coherence and conciseness (Mani, 2001). In the following, we describe for each of the four considered tasks (*i.e.*, NMT, text summarization, image captioning and data2text generation) the most used metrics.

Metrics for NMT. Most of the metrics commonly used in NMT rely on comparing surface form (*e.g.*, word, subword, n-gram overlap and edit based distances (Levenshtein, 1966)) between text and candidates. Perhaps the most popular metrics are the ones used for WMT shared tasks (Mathur et al., 2020; Ma et al., 2019, 2018; Bojar et al., 2017b) which include SENTBLEU, BLEU (Papineni et al., 2002), CHARACTER (Wang et al., 2016), COMET (Rei et al., 2020), YISI (Lo et al., 2018), MEE (Mukherjee et al., 2020), EED (Stanchev et al., 2019), CHRF (Popović, 2015, 2017), ESIM (Chen et al., 2016), PRISM (Thompson and Post, 2020) to only mention a few among others. A new family of metrics based on pretrained transformers (*i.e.*, BertScore, MoverScore) has recently emerged with very good performance in NMT, incorporating deeper semantic information through contextualized representations.

Metrics for summarization. Designing better summarization metrics is an active area of research (Scialom et al., 2021) and many of these metrics can be further optimized to produce better summaries (Böhm et al., 2019). Popular metrics include machine translation metrics (*i.e.*, CHRF, BLEU, METEOR (Banerjee and Lavie, 2005; Guo and Hu, 2019; Denkowski and Lavie, 2014), BertScore, MoverScore or SentenceMover (Clark et al., 2019)), ROUGE (Lin, 2004; Ganesan, 2018) or data statistics (*e.g.*, density and compression ratio) (Grusky et al., 2018).

Metrics for data2text. Data2text generation aims at generating text from structured data (Kim and Mooney, 2010; Chen and Mooney, 2008; Wiseman et al., 2017). In the present work, we focus on the WebNLG 2020 challenge (Perez-Beltrachini et al., 2016; Gardent et al., 2017) which ranks the system using five automatic metrics: BLEU, METEOR, BERTScore, TER and CHRF++.

Metrics for image captioning. Task specific metrics for image captioning include CIDEr (Vedantam et al., 2015) that rely on n-grams, LEIC (Cui et al., 2018) using scene graph similarity and pretrained metrics such as SPICE (Anderson et al., 2016). In

recent work by (Zhang et al., 2019; Zhao et al., 2019b), these metrics are compared with NMT specific metrics (e.g., BLEU and METEOR).

3 Background on Optimal Transport

The Wasserstein distance (*i.e.*, Earth Mover Distance) which arises from the idea of optimal transport provides a way to measure dissimilarities between two probability distributions. Due to its appealing geometric properties, it has found many applications in machine learning such as generative models (Arjovsky et al., 2017; Tolstikhin et al., 2018; Gulrajani et al., 2017), domain adaptation (Courty et al., 2017), clustering (Ho et al., 2017; Ye et al., 2017), adversarial examples (Wong et al., 2019), robustness (Staerman et al., 2021) or NLP (Kusner et al., 2015; Zhao et al., 2019b; Singh et al., 2020). First designed as an optimal transport optimization problem, it relies on minimizing a transport cost between points drawn from all possible coupling measures. Its ability to take into account the underlying geometry of the space as well as capture information from distributions with non-overlapping supports makes it a powerful alternative to several dissimilarity measures such as the family of f-Divergences.

Wasserstein distance. Let $\mathcal{M}_+^1(\mathbb{R}^d)$ denote the space of all probability distributions defined on $\mathbb{R}^d$ with $d \in \mathbb{N}^*$. The Wasserstein distance between two arbitrary measures $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ is defined through the resolution of the Monge-Kantorovitch mass transportation problem (Villani, 2003; Peyré and Cuturi, 2019):

$$\mathcal{W}(\mu, \nu) = \min_{\pi \in \mathcal{U}(\mu,\nu)} \left(\int_{\mathcal{X} \times \mathcal{Y}} ||x - y||^p d\pi(x,y) \right)^{1/p} \quad (1)$$

where $\mathcal{U}(\mu, \nu) = \{ \pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : \int \pi(x,y)dy = \mu(x); \int \pi(x,y)dx = \nu(y) \}$ is the set of joint probability distributions with marginals μ and ν. In the remainder of this paper, we focus on the Wasserstein distance associated with the quadratic cost, *i.e.*, $p = 2$. Thus, the Wasserstein distance aims to find the best possible way to transfer the probability mass from μ to ν while minimizing the transportation cost defined by the euclidian distance.

Wasserstein barycenters. Because optimal transport is based on mass displacement, it also defines an interesting way to interpolate between several input measures. The Wasserstein barycenter,

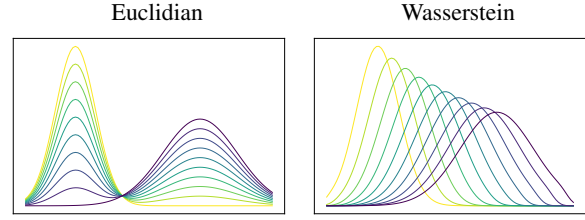Euclidian   Wasserstein

Figure 1: Euclidian (left) and Wasserstein (right) interpolation between densities of two Gaussian distributions.

first introduced and studied in (Agueh and Carlier, 2011), defines an interpolation measure between several probability distributions. The main asset of Wasserstein barycenters is to take into account the geometry of the space where input measures live in (cf. Fig. 1). Given N probability distributions: $\mu_1, \ldots, \mu_N \in \mathcal{M}_+^1(\mathbb{R}^d)$ and weights $(\alpha_1, \ldots, \alpha_N) \in \mathbb{R}_+$, the Wasserstein barycenter optimization problem of these distributions w.r.t. the weights is defined as:

$$\mu = \underset{\mu \in \mathcal{M}_+^1(\mathbb{R}^d)}{\operatorname{argmin}} \sum_{i=1}^N \alpha_i \mathcal{W}(\mu_i, \mu), \quad (2)$$

where the support of μ may be unknown. Equation 2 defines a weighted average in the Wasserstein space. To make it computationally tractable, the measure μ is often constrained to be a discrete measure with free (Cuturi and Doucet, 2014; Álvarez Esteban et al., 2016; Cuturi and Peyré, 2016; Luise et al., 2019) or fixed support (Benamou et al., 2015; Dvurechenskii et al., 2018; Lin et al., 2020; Janati et al., 2020). For the purpose of our approach, we focus on free support barycenter with fixed weights in the following.

4 `BaryScore` Metric

The construction of automatic metrics usually relies on two paradigms depending on the availability of a reference sentence for each candidate (Specia et al., 2010). Here and throughout the paper, we assume that at least one reference is available for each candidate. Denote by $C = \{\omega_1^c, \ldots, \omega_{n_c}^c\}$ the candidate and $R = \{\omega_1^r, \ldots, \omega_{n_r}^r\}$ the reference composed of n_c and n_r words, respectively. Our goal is to design a metric $m : (C, R) \mapsto m(C, R) \in \mathbb{R}_+$ such that the closer to zero the better candidate is.

Algorithm. Our metric m, named `Baryscore`, can be summarized in two steps:

(i) Find the Wasserstein barycentric distributions of contextual encoder layers for C and R;
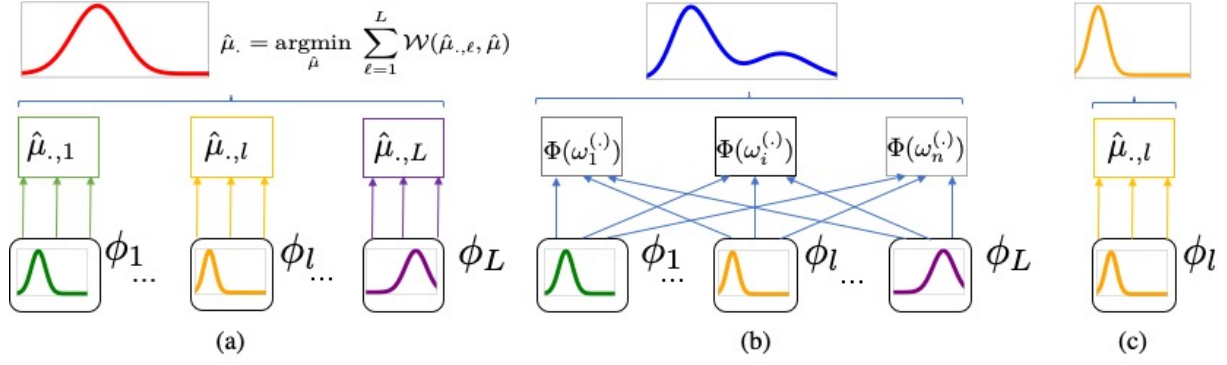
Figure 2: Schema of the different layers aggregation step: (a) Wasserstein barycenter (as used in `BaryScore`), (b) Power mean (as used in MoverScore), (c) best layer selection (case of BertScore). $\hat{\mu}_{.,\ell}$ $\ell \in [1, L]$ stands for either the empirical distribution of the reference R or the candidate C text. Similarly, for words denoted by $\omega_i^{\cdot}$.

(ii) Evaluate these barycentric distributions using the Wasserstein distance.

Wasserstein barycenters. Assume that a contextual encoder, (*e.g.* BERT and ELMo), is composed of L layers, *i.e.*, $\phi_1, \ldots, \phi_L$ functions that map a candidate text C and a reference text R respectively into $\phi_\ell(C) \in \mathbb{R}^{n_c \times d}$ and $\phi_\ell(R) \in \mathbb{R}^{n_r \times d}$, for every $1 \leq \ell \leq L$. In our approach, we consider the discrete probability distributions induced by $\phi_\ell(C)$ and $\phi_\ell(R)$, where $\phi_\ell(C)_i$ and $\phi_\ell(R)_j$ represent the embedding of the i-th token and j-th token of the candidate and reference text, respectively. Precisely, $2L$ empirical measures are constructed from these layers functions such that $\hat{\mu}_{C,\ell} = \sum_{i=1}^{n_c} \alpha_i \delta_{\phi_\ell(\omega_i^c)}$ and $\hat{\mu}_{R,\ell} = \sum_{j=1}^{n_r} \beta_j \delta_{\phi_\ell(\omega_j^r)}$, where $\alpha = \{\alpha_1, \ldots, \alpha_{n_c}\}$ and $\beta = \{\beta_1, \ldots, \beta_{n_r}\}$ are the vector of inverse document frequencies of each word ω_i of C and R, respectively, and δ_x is the dirac mass at point x. Further, Wasserstein barycenters (see Eq. (2)) are computed on the candidate C and the reference R leading to two barycentric embedding measures $\hat{\mu}_C$ and $\hat{\mu}_R$ with fixed sizes n_c and n_r, respectively. Considering weights of the barycentric measures as uniform, as for the layers weights, the optimization problem is equivalent to find locations $\{x_i^c\}_{i=1}^{n_c}$ such that:

$$\underset{(x_1^c, \ldots, x_{n_c}^c) \in \mathbb{R}^d}{\text{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{c,\ell}, \hat{\mu}_c) \quad (3)$$

with $\hat{\mu}_c = \frac{1}{n_c} \sum_{i=1}^{n_c} \delta_{x_i^c}$. The previous formulation is similar for the reference text, replacing C by R and α by β in notations. The final embedding, denoted by Φ, is then considered as the locations of the barycentric measures, *i.e.*, $\Phi(C) = \{x_1^c, \ldots, x_{n_c}^c\}$ and $\Phi(R) = \{x_1^r, \ldots, x_{n_r}^r\}$.

Computing the Wasserstein distance. The last step of our approach is then to evaluate discrete measures induced by the final embeddings, *i.e.*, the candidate and reference barycentric measures $\hat{\mu}_C$ and $\hat{\mu}_R$, using the Wasserstein distance, leading to the `Baryscore` given by $m(C, R) = \mathcal{W}(\hat{\mu}_C, \hat{\mu}_R)$. This step boils down to computing the WMD and is similar to the final step in MoverScore or SentenceMover. The entire procedure is summarized in Algorithm 1.

Algorithm 1 `BaryScore`

INPUT: $C = \{\omega_1^c, \ldots, \omega_{n_c}^c\}$, $R = \{\omega_1^r, \ldots, \omega_{n_r}^r\}$, $(\phi_1, \ldots, \phi_L)$ pre-trained layers from BERT or ELMo.

Compute layers embeddings:
$\phi_\ell(C)$ and $\phi_\ell(R)$ for every $1 \leq \ell \leq L$.

Compute measures: $\{\hat{\mu}_{C,\ell}, \hat{\mu}_{R,\ell}\}_{\ell=1}^{L}$.

Compute Wasserstein barycenters:

$$\hat{\mu}_C = \underset{\hat{\mu}}{\text{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{C,\ell}, \hat{\mu}),$$

$$\hat{\mu}_R = \underset{\hat{\mu}}{\text{argmin}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{R,\ell}. \hat{\mu}),$$

OUTPUT: $\mathcal{W}(\hat{\mu}_R, \hat{\mu}_C)$.

Parameters of BaryScore metric. Our metric is dependant on the choice of the continuous representations (*e.g.*, BERT, ELMo and Roberta) and then its performance will be influenced by the choice of the model. As it is common in concurrent work (Zhang et al., 2019; Zhao et al., 2019b; Clark et al., 2019) all the results in the paper are obtained with

10453

one single model: namely BERT-base-uncased. Additionally, we report results obtained with the BERT fine-tuned on NLI release in (Zhao et al., 2019b), this model is referred as `BaryScore⁺` in the following. In contrast to previous work (*e.g* SentenceMover) that integrates a preprocessing step by removing the stopwords based on a static list, we keep all words. Also, our framework provides a natural way to exploit all available layers of the model while previous work relies on a specific subset of them (*e.g.*, MoverScore and BertScore). We believe this strengthens the robustness of our approach.

Comparison with the Moverscore. Following the footstep of (Zhang et al., 2019), the Moverscore (Zhao et al., 2019b) applied optimal transport to the output of Contextualized Encoders (CE) such as BERT or ELMo. Precisely, let's assume that a CE is composed of L layers, the Moverscore's context representation is given for each word ω_j by $\Phi(\omega_j) = T(\phi_1(\omega_j), \ldots, \phi_L(\omega_j))$, where the transformation T is either power means (Rücklé et al., 2018a) or aggregation routines depicted in (Zhao et al., 2018, 2019a). The score is then defined by the Wasserstein distance between the empirical distributions given by $\Phi(C)$ and $\Phi(R)$.

The main weakness of this approach is the aggregation step. Taking into account the role of the underlying geometry of the probability distribution as well as the interpretability of the transportation flow are high benefits of Optimal transport. However, performing Wasserstein distance after applying power means, *i.e.*, an aggregation in an euclidian space (see *e.g.*, Figure 1), does not allow a proper evaluation of the geometry induced by the CE layers in the Wasserstein space. Indeed, Moverscore evaluates a distorted geometry inducing wrong interpretability of the transportation flow. The advantage of exploiting Wasserstein barycenter over euclidean aggregation relies on rehabilitating this geometry, as shown in Section 6.

5 Experimental Settings

In this section, we present our evaluation methods as well as the various dataset used to benchmark our metric.

Extension of notations. In the previous section, we have only considered a candidate and a reference sentence. In order to evaluate and compare different metrics, we need to extend the previous notations to include the system that generates each sentence. To this end, we will assume that we have

a dataset: $\mathcal{D} = \{R_i, \{C_i^s, h(C_i^s)\}_{s=1}^S\}_{i=1}^N$, where C_i^j is the i-th text generated by the j-th system, $h(C_i^j)$ is human score assign to C_i^j and R_i the reference text associated to C_i^j; N is the number of available texts; S the number of different systems.

5.1 Evaluating automatic evaluation of NLG

The quality of the evaluation metric is measured by its correlation with the human judgment (Chatzikoumi, 2020; Specia et al., 2010; Koehn, 2009; Banerjee and Lavie, 2005). Three correlation measures can be considered: Pearson (Leusch et al., 2003), Spearman (Melamed et al., 2003) or Kendall (Kendall, 1938). In addition, two different levels of granularity are considered to compute theses correlation coefficients.

System level correlation. These can be considered when assessing the discrimination capability between two systems. This level of correlation tries to answer the question: "*Can the metric be used to compare the performance of two systems?*". Formally, the system level correlation K_{sys} measures the quality of a metric m defined as:

$$K_{sys} = K(M^{sy}, H^{sy}), \tag{4}$$

$$M^{sy} = \left[\frac{1}{N}\sum_{i=1}^N m(R_i, C_i^1), \cdots, \frac{1}{N}\sum_{i=1}^n m(R_i, C_i^s)\right],$$

$$H^{sy} = \left[\frac{1}{N}\sum_{i=1}^N h(C_i^1), \cdots, \frac{1}{N}\sum_{i=1}^N h(C_i^S)\right],$$

where K is the considered correlation coefficient.

Text level correlation. This is computed to evaluate the ability of a metric to measure the semantic equivalence between a candidate and a reference sentence. Such a level of correlation aims at providing an answer to the question: "*Can the metric be used as a loss or reward of a system?*". By introducing similar notations to those in Equation 4, we obtain the text level correlation K_{text}:

$$K_{text} = \frac{1}{N}\sum_{i=1}^N K(M_i^{text}, H_i^{text}), \tag{5}$$

$$M_i^{text} = \left[m(R_i, C_i^1), \cdots, m(R_i, C_i^S)\right],$$

$$H_i^{text} = \left[h(C_i^1), \cdots, h(C_i^S)\right].$$

Significance testing. To ensure that observed improvement is statistically significant we follow common consensus (Deutsch et al., 2021; Graham, 2015; Graham et al., 2015; Graham and Baldwin,

2014) in NMT and rely on William test (Steiger, 1980) as considered observations are correlated[3].

5.2 Choice of datasets

We motivate our choice of datasets for each tasks.
Translation. Multiple translation datasets are available from the WMT translation shared tasks (Bojar et al., 2014, 2015, 2016, 2017a). Keeping in mind the work by (Card et al., 2020) stressing the importance of the size of the dataset, we follow (Zhang et al., 2019) and choose to work with WMT16 and additionally report results on WMT15 that both offer over 500 sentences per language (in contrast to new versions (Bojar et al., 2018; Barrault et al., 2019, 2020) that rely on a lower number–around 50–of annotated texts).

Summarization. Several datasets have been introduced to compare metrics for summarization. Classical choices include MSR Abstractive Text Compression dataset (Toutanova et al., 2016), TAC datasets (Dang and Owczarzak, 2008; McNamee and Dang, 2009) or on news summarization from CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016). TAC datasets and contains flaws (Rankel et al., 2013; Peyrard, 2019b; Bhandari et al., 2020), thus in this work we rely on the CNN introduced in (Bhandari et al., 2020). It is composed of 11,490 summaries comming from 11 extractive systems (Lewis et al., 2019; Yoon et al., 2020; Raffel et al., 2019; Gehrmann et al., 2018; Dong et al., 2019; Liu and Lapata, 2019; Chen and Bansal, 2018; See et al., 2017) and 14 abstractive systems (Zhong et al., 2020; Wang et al., 2020; Zhong et al., 2019; Liu and Lapata, 2019; Zhou et al., 2018; Narayan et al., 2018; Dong et al., 2019; Kedzie et al., 2018; Zhou et al., 2018).

Data2Text. In contrast to previous work that rely on old task-oriented dialogue datasets (*i.e.*, BAGEL (Mairesse et al., 2010), SFHOTEL (Wen et al., 2015)), we focus on the WebNLG challenge (Perez-Beltrachini et al., 2016; Gardent et al., 2017; Ferreira et al., 2018, 2020) as sentence available in this challenge is more representative of progress of NLG[4]. The data provides from 15 systems relying on either symbolic or neural approaches. In this challenge, multiple evaluation criteria are used to assess the quality of the text. Due to the space limitations, we compute correlation on the three

---

[3]An example code is provided by the authors at `https://github.com/ygraham/nlp-williams`

[4]System description and performance are available at `https://webnlg-challenge.loria.fr/`
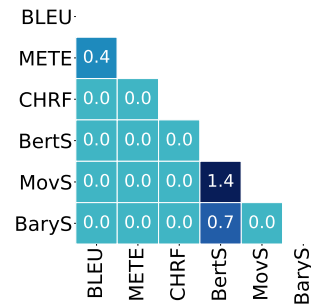

Figure 3: Significance testing on de-en for WMT16. In the matrix is reported the p-value of the William Test in percent.

following criteria: (1) *data coverage* which measures if the generated text contains all the available information present in the input data, (2) *relevance* which characterizes if the generated text is solely composed with information available in the input, (3) *correctness* which measures if all input information is both correct and adequately introduced.
Image captioning. To evaluate metrics on image captioning, there is a consensus to exploit COCO datasets (Lin et al., 2014) and to compute correlation at the system level. On average each image has five reference captions. Each of the 12 systems is evaluated on 5 criteria. Following (Anderson et al., 2016; Zhao et al., 2019b; Zhang et al., 2019), we only compute correlation on M1, M2 which are related to the overall quality of the caption.

6 Numerical Results

In this section, we study the performance of `BaryScore` on the four aforementioned tasks.

6.1 Translation

Overall results. Table 1 and Table 2 gather correlation to human judgments on WMT15 and WMT16. We conduct a statistical analysis to ensure that the observed improvements are statistically meaningful (see Figure 3).

We observe that `BaryScore`[+] is the best performing metric on both datasets for all languages. Similarly to (Zhao et al., 2019b), we observe an improvement when using their pretrained version of BERT on MNLI (Wang et al., 2018). By comparing the best performance achieved by `BaryScore` compared to `MoverScore`, we hypothesize that Wasserstein barycenter preserves more geometric properties of the information learnt by BERT.
Correlation analysis. Figure 4 reports the inter-correlation across metrics according to the Kendall

	cs - en			de - en			ru - en			fi - en		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
BaryS⁺	**75.9**	**75.2**	**56.9**	**75.8**	**74.5**	**56.2**	**77.6**	**75.0**	**56.7**	**79.9**	**78.7**	**59.7**
BaryS	74.2	73.8	54.9	74.1	72.2	53.1	73.7	70.6	53.2	76.6	74.5	56.3
BertS-F	74.3	73.5	54.3	72.2	70.7	52.9	74.0	70.5	52.5	74.7	72.5	54.1
MoverS⁺	71.0	71.1	52.0	71.1	68.2	50.3	67.3	64.7	47.3	72.2	72.0	52.9
MoverS	68.8	70.1	52.0	71.8	69.4	50.3	68.6	65.5	46.9	70.0	70.0	52.3
MET	56.4	57.1	40.9	60.1	61.4	61.5	58.9	59.6	42.4	58.8	59.0	42.9
BLEU	44.5	44.5	31.1	53.6	48.0	48.1	53.4	49.0	34.1	46.5	41.8	29.3
CHRF	26.0	21.6	15.8	29.5	28.9	28.9	32.5	32.9	23.4	30.2	26.5	19.2

Table 1: Absolute correlations between metric prediction and text level human judgement on 4 pairs of WMT15.

	cs - en			de - en			ru - en			fi - en			ro - en			tr - en		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
BaryS⁺	**76.6**	**76.2**	**57.5**	**68.5**	**67.7**	**50.0**	**69.4**	**68.3**	**51.3**	**70.2**	**69.5**	**50.9**	**74.3**	**73.0**	**54.5**	**73.8**	**70.5**	**52.4**
BaryS	74.2	74.3	56.3	64.6	64.2	47.9	67.5	66.4	48.1	67.1	66.4	48.3	72.5	71.4	52.9	69.3	67.1	51.4
BertS-F	74.1	73.6	55.49	65.3	64.6	46.3	65.1	64.6	46.9	65.4	64.1	47.0	70.2	67.6	49.5	70.7	67.1	49.0
MoverS⁺	70.7	70.4	53.4	62.4	60.7	44.8	64.0	62.2	45.2	64.5	62.6	45.8	66.4	66.0	48.6	66.3	60.7	44.9
MoverS	67.4	69.5	52.6	60.9	59.1	44.2	64.4	62.8	44.8	63.1	62.2	45.1	64.2	65.4	48.2	66.1	64.0	43.7
MET	64.5	67.2	49.2	51.6	50.0	35.3	54.8	57.5	41.2	53.9	52.4	37.3	58.6	59.4	42.1	61.8	59.1	42.4
BLEU	53.8	52.3	36.2	45.3	40.8	28.2	46.3	43.8	30.2	39.9	37.7	26.3	47.0	43.2	30.2	47.1	43.9	30.4
CHRF	24.4	26.8	19.2	34.2	33.8	24.3	28.7	29.8	21.2	14.4	15.3	11.2	16.0	12.6	9.0	26.5	16.2	12.4

Table 2: Absolute correlations between metric prediction and text level human judgement on 6 pairs of WMT16.
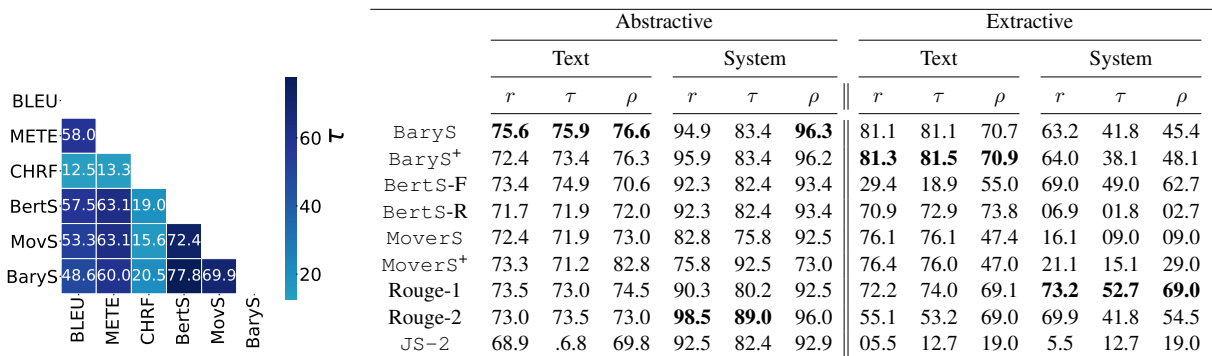

Figure 4: de-en WMT16 Kendall correlation between predictions made by different metrics.

	Abstractive						Extractive					
	Text			System			Text			System		
	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ
BaryS	**75.6**	**75.9**	**76.6**	94.9	83.4	**96.3**	81.1	81.1	70.7	63.2	41.8	45.4
BaryS⁺	72.4	73.4	76.3	95.9	83.4	96.2	**81.3**	**81.5**	**70.9**	64.0	38.1	48.1
BertS-F	73.4	74.9	70.6	92.3	82.4	93.4	29.4	18.9	55.0	69.0	49.0	62.7
BertS-R	71.7	71.9	72.0	92.3	82.4	93.4	70.9	72.9	73.8	06.9	01.8	02.7
MoverS	72.4	71.9	73.0	82.8	75.8	92.5	76.1	76.1	47.4	16.1	09.0	09.0
MoverS⁺	73.3	71.2	82.8	75.8	92.5	73.0	76.4	76.0	47.0	21.1	15.1	29.0
Rouge-1	73.5	73.0	74.5	90.3	80.2	92.5	72.2	74.0	69.1	**73.2**	**52.7**	**69.0**
Rouge-2	73.0	73.5	73.0	**98.5**	**89.0**	96.0	55.1	53.2	69.0	69.9	41.8	54.5
JS-2	68.9	.6.8	69.8	92.5	82.4	92.9	05.5	12.7	19.0	5.5	12.7	19.0

Figure 5: Absolute correlation coefficients (as measured using Pearson (r), Spearman (ρ) and Kendall (τ) coefficient) between different metrics on text summarization.

τ. We observe that the metrics based on BERT (*e.g* BertScore, MoverScore and BaryScore) obtain medium-high correlation demonstrating that both the aggregation mechanism (*e.g.*, one layer selection, power mean or Wasserstein barycenter) as well as the choice of similarity metric (*e.g.*, cosine similarity, Wasserstein distance) affects the ranking of the predictions.

Takeaways: Overall BaryScore is particularly suitable to compare two examples and thus could be used as an alternative to the standard cross-entropy loss to train NMT systems. On the other hand, our implementation made based on POT (Flamary and Courty, 2017) makes the speed comparable with MoverScore. We are able to process over 180 sentence pairs per second with BaryScore compared to 195 sentence pairs per second with MoverScore on an NVIDIA-V100 GPU.

6.2 Summarization

Figure 5 reports results on the summarization task. We are able to reproduce the performance reported in the original paper (Bhandari et al., 2020). Contrarily to MT, we observe that there is no metric that can outperform all others on all correlation measurements. We can also notice that the improvement induced by the BERT fine-tuned on MNLI is not observed on this dataset for both BaryScore and MoverScore.

	M1	M2
BaryS$^+$	85.6	83.9
BaryS	85.2	82.6
MoveS$^+$	83.1	82.6
MoveS	78.1	82.1
BertS-P	83.0	82.1
BertS-R	80.1	75.2
BertS-F	79.1	81.1
BLEU	58.3	60.8
METEOR	60.3	59.9
SPICE*	75.9	75.0
CIDER*	43.8	44.0
LEIC*	**93.9**	**94.9**

Figure 6: System level Pearson correlation with human judgement on the MSCOCO dataset.

	Correctness			Data Coverage			Relevance		
Metric	r	ρ	τ	r	ρ	τ	r	ρ	τ
Correct	100.0	100.0	100.0	97.6	85.2	73.3	99.1	89.7	75.0
DataC	85.2	97.6	73.3	100.0	100.0	100.0	96.0	93.8	81.6
Relev	89.7	99.1	75.0	96.0	93.8	81.6	100.0	100.0	100.0
BaryS	**91.7**	**90.0**	**78.3**	**87.8**	78.2	61.6	**89.4**	**82.6**	70.0
BaryS$^+$	90.5	89.5	76.6	87.7	**85.0**	**70.0**	89.2	86.4	71.6
BertS	85.5	83.4	73.3	74.7	68.2	53.3	83.3	79.4	65.0
MoverS	84.1	84.1	73.3	78.7	66.2	53.3	82.1	77.4	65.0
BLEU	77.6	66.3	60.0	55.7	50.2	36.6	63.0	65.2	51.6
R-1	80.6	65.0	65.0	76.5	76.3	60.3	64.3	69.2	56.7
R-2	73.6	63.3	58.3	54.7	43.1	35.0	62.0	60.8	46.7
R-WE	60.9	73.4	60.0	40.2	58.2	40.1	49.9	64.1	48.3
METEOR	86.5	66.3	70.0	77.3	50.2	46.6	82.1	65.2	58.6
TER	79.6	78.3	58.0	69.7	58.2	38.0	75.0	70.2	**77.6**

Figure 7: Correlation at the system level with human judgement along five different axis: correctness, data coverage, fluency, relevance and text structure for the WebNLG task. Overall best result is bolted.

Consistency and robustness of `BaryScore`. In contrast to what is observed on abstractive systems, we observe a strong inconsistency in the behavior of the previous metrics based on BERT for extractive systems. Indeed, at the text level BertS-R, MoverScore and MoverScore$^+$ achieve good medium/high correlation whereas at the system level the achieved correlation collapses (correlation scores below 20 points). BertS-F, on the contrary, under-performs at the text level for extractive systems but achieves competitive performance with Rouge (the best performing metric) at the system level. We observe that using Wasserstein barycenter is a better way to aggregate the layer and provides better robustness as it alleviates the aforementioned problem. Indeed, the performance achieved by `BaryScore` is competitive at both the text and system levels.

Takeaways: Overall, the two versions of `BaryScore` are among the best performing metrics, outperform current BERT based metrics on 3/4 configurations, and achieve consistent performance on the 4th configuration. The consistent behavior of `BaryScore` demonstrates the validity of our approach for summarization. Whereas, a simpler and lighter alternative to `BaryScore`, as well as other BERT-based metrics to compare systems on summarization, remains the ROUGE score for 3/4 configurations.

6.3 Data2Text

Figure 7 reports results on data2text task using the WebNLG2020 data. To the best of our knowledge this is one of the first study using this dataset.

Figure 7 shows a strong correlation between the three evaluation dimensions with correlation r and ρ higher than 90. We observe that `BaryScore` consistently metrics based on BERT and achieves best results on 6/9 configurations for `BaryScore` and 2/9 for `BaryScore`$^+$.

6.4 Image Captioning

We follow (Zhang et al., 2019; Zhao et al., 2019b) and report in Figure 6[5] Pearson correlation coefficients between prediction and system level judgment. Although we were unable to reproduce exactly the results by (Zhao et al., 2019b), we obtain comparable numbers and similar orderings.

Takeaways: `BaryScore` outperforms current metrics except for `LEIC` that rely on information extracted from both image and text. These results validate the use of `BaryScore` to compare the performance of image captioning systems.

7 Summary and Concluding Remarks

In this paper, we present a metric named `BaryScore` which relies on optimal transport and solves the geometric discrepancies present in existing metrics that use contextualized embedding with WMD. The present work is carried out in the context of NLG but it introduces a generic theoretically-grounded framework that could be extended to other NLP studies. In particular, it illustrates applications of Wasserstein barycenters to combine the different views offered by different layers of a deep neural network. Specifically,

---

[5]Results with * are reported from (Zhao et al., 2019b)

10457

futur work includes testing Wasserstein Barycenters in a multimodal setting (Garcia et al., 2019; Colombo et al., 2021a), for classification (*e.g.* emotion (Witon et al., 2018), dialog act (Chapuis et al., 2020; Colombo et al., 2020; Chapuis et al., 2021), stance (Dinkar et al., 2020)) and controlling style in NLG (Jalalzai et al., 2020; Colombo et al., 2019, 2021b) .

8 Acknowledgment

References

Martial Agueh and Guillaume Carlier. 2011. Siam journal on mathematical analysis. *Barycenters in the Wasserstein Space*, 43(2):904–924.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. 2015. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138.

Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*.

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. *arXiv preprint arXiv:1909.01214*.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017a. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Trans-

lation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017b. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. *arXiv preprint arXiv:2010.06595*.

Emile Chapuis, Pierre Colombo, Matthieu Labeau, and Chloé Clave. 2021. Code-switched inspired losses for generic spoken dialog representations. *CoRR*, abs/2108.12465.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloe Clavel. 2020. Hierarchical Pre-Training for Sequence Labelling in Spoken Dialog. *arXiv preprint arXiv:2009.11152*.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

David L Chen and Raymond J Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

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.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. 2021a. Improving multimodal fusion via mutual dependency maximisation. *arXiv preprint arXiv:2109.00922*.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. *arXiv preprint arXiv:2002.08801*.

Pierre Colombo, Chloé Clavel, and Pablo Piantanida. 2021b. A novel estimator of mutual information for learning to disentangle textual representations. *CoRR*, abs/2105.02685.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-Driven Dialog Generation. 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 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicolas Courty, Remy Flamary, Devis Tuia, and Alain Rakotomamonjy. 2017. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865.

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.

Marco Cuturi and Arnaud Doucet. 2014. Fast computation of wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 685–693.

Marco Cuturi and Gabriel Peyré. 2016. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. In *TAC*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *arXiv preprint arXiv:2104.00054*.

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

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. *arXiv preprint arXiv:2009.11340*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Pavel Dvurechenskii, Darina Dvinskikh, Alexander Gasnikov, Cesar Uribe, and Angelia Nedich. 2018. Decentralize and randomize: Faster algorithm for wasserstein barycenters. In *Advances in Neural Information Processing Systems*, volume 31.

Pedro C. Álvarez Esteban, E. del Barrio, J.A. Cuesta-Albertos, and C. Matrán. 2016. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762.

Thiago Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional webnlg+ shared task overview and evaluation results (webnlg+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*.

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the webnlg corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176.

Rémi Flamary and Nicolas Courty. 2017. Pot python optimal transport library. *GitHub: https://github.com/rflamary/POT*, page 144.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

Alexandre Garcia, Pierre Colombo, Florence d'Alché Buc, Slim Essid, and Chloé Clavel. 2019. From the Token to the Review: A Hierarchical Multimodal Approach to Opinion Mining. 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 5539–5548, Hong Kong, China. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planning. In *55th annual meeting of the Association for Computational Linguistics (ACL)*.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Proceedings of the 30rd Conference on Neural Information Processing Systems (NeurIPS)*.

Yinuo Guo and Junfeng Hu. 2019. Meteor++ 2.0: Adopt syntactic level paraphrase knowledge into machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 501–506, Florence, Italy. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *arXiv preprint arXiv:1506.03340*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.

Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. 2017. Multilevel clustering via Wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1501–1509.

Eduard H Hovy. 1999. Toward finely differentiated evaluation metrics for machine translation. In *Proceedings of the EAGLES Workshop on Standards and Evaluation Pisa, Italy, 1999*.

Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. *arXiv preprint arXiv:2003.11593*.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. 2020. Debiased Sinkhorn barycenters. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 4692–4701.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. *arXiv preprint arXiv:1810.12343*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Joohyun Kim and Raymond Mooney. 2010. Generative alignment and semantic parsing for learning from ambiguous supervision. In *Coling 2010: Posters*, pages 543–551.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Gregor Leusch, Nicola Ueffing, Hermann Ney, et al. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of Mt Summit IX*, pages 240–247.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Quentin Lhoest, Patrick von Platen, Thomas Wolf, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Suraj Patil, Lewis Tunstall, Mariama Drame, Julien Chaumond, Julien Plu, Joe Davison, Simon Brandeis, Victor Sanh, Teven Le Scao, Kevin Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, and François Lagunas. 2021. huggingface/datasets: 1.9.0.

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.

Tianyi Lin, Nhat Ho, Xi Chen, Marco Cuturi, and Michael Jordan. 2020. Fixed-support wasserstein barycenters: Computational hardness and fast algorithm. In *Advances in Neural Information Processing Systems*, volume 33, pages 5368–5380.

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.

Lucian Vlad Lita, Monica Rogati, and Alon Lavie. 2005. Blanc: Learning evaluation metrics for mt. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 740–747.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Chi-kiu Lo, Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell. 2018. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The NRC supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 908–916, Belgium, Brussels. Association for Computational Linguistics.

Chi-kiu Lo and Dekai Wu. 2011. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 220–229.

Ryan Lowe, Iulian V Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. On the evaluation of dialogue systems with next utterance classification. *arXiv preprint arXiv:1605.05414*.

Giulia Luise, Saverio Salzo, Massimiliano Pontil, and Carlo Ciliberto. 2019. Sinkhorn barycenters with free support via frank-wolfe algorithm. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the wmt18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the third conference on machine translation: shared task papers*, pages 671–688.

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.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

François Mairesse, Milica Gasic, Filip Jurcicek, 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.

Inderjeet Mani. 2001. *Automatic summarization*, volume 3. John Benjamins Publishing.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, volume 17, pages 111–113.

I Dan Melamed, Ryan Green, and Joseph Turian. 2003. Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, pages 61–63.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Ananva Mukherjee, Hema Ala, Manish Shrivastava, and Dipti Misra Sharma. 2020. Mee: An automatic metric for evaluation using embeddings for machine translation. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 292–299. IEEE.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Laura Perez-Beltrachini, Rania Sayed, and Claire Gardent. 2016. Building rdf content for data-to-text generation. In *The 26th International Conference on Computational Linguistics (COLING 2016)*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of 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.

Gabriel Peyré and Marco Cuturi. 2019. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Peter A Rankel, John Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

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

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018b. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2020. A survey of evaluation metrics used for nlg systems. *arXiv preprint arXiv:2008.12009*.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Safeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

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: Shared Task Papers*, pages 751–758.

Sidak Pal Singh, Andreas Hug, Aymeric Dieuleveut, and Martin Jaggi. 2020. Context mover's distance barycenters: Optimal transport of contexts for building representations.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Guillaume Staerman, Pierre Laforgue, Pavlo Mozharovskyi, and Florence d'Alché Buc. 2021. When ot meets mom: Robust estimation of wasserstein distance. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 136–144.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520, Florence, Italy. Association for Computational Linguistics.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. *arXiv preprint arXiv:2004.14564*.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. 2018. Wasserstein autoencoders. In *6th International Conference on Learning Representations (ICLR 2018)*.

Kristina Toutanova, Chris Brockett, Ke M Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Cedric Villani. 2003. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, New York.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. *arXiv preprint arXiv:1909.01380*.

Alex Wang, Amanpreet 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*.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510.

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*.

John S White, Theresa A O'Connell, and Francis E O'Mara. 1994. The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Wojciech Witon, Pierre Colombo, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at IEST 2018: Predicting emotions using an ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 248–253. Association for Computational Linguistics.

Eric Wong, Frank Schmidt, and Zico Kolter. 2019. Wasserstein adversarial examples via projected Sinkhorn iterations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6808–6817.

Jianbo Ye, Panruo Wu, James Z. Wang, and Jia Li. 2017. Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332.

Wonjin Yoon, Yoon Sun Yeo, Minbyul Jeong, Bong-Jun Yi, and Jaewoo Kang. 2020. Learning by semantic similarity makes abstractive summarization better. *arXiv preprint arXiv:2002.07767*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Haiyun Peng, Steffen Eger, Erik Cambria, and Min Yang. 2019a. 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.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019b. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

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 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3110–3119.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. *arXiv preprint arXiv:1907.03491*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. *arXiv preprint arXiv:1807.02305*.

9 Appendix

We gather additional experimental results. In particular, a statistical analysis on WMT16 and WebNLG2020.

9.1 Statistical analysis of WMT16

We report on Figure 11 the correlation coefficients the inter-correlation across metrics on WMT16.

9.2 Statistical analysis of data2text

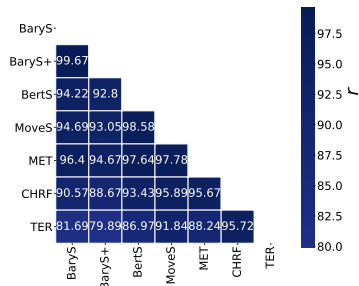We report in Figure 15 the results of the William test on data2text generation. (Lhoest et al., 2021)
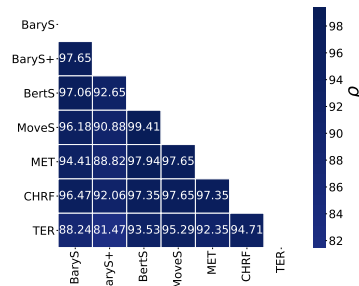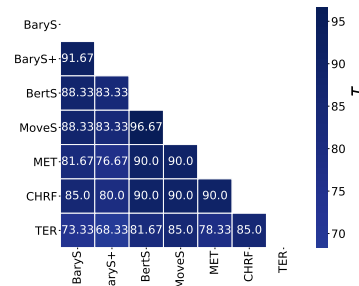
Figure 8: Pearson r

Figure 9: Spearman ρ

Figure 10: Kendall τ

Figure 11: Correlation coefficients measuring correlation between the prediction made by different metrics based on BERT on WMT16 (de-en).
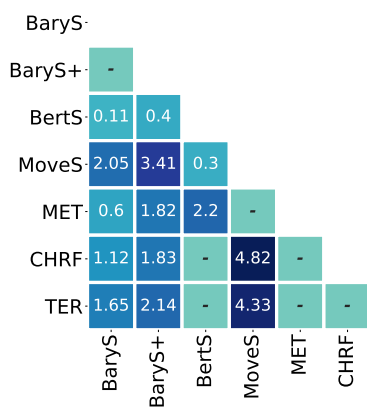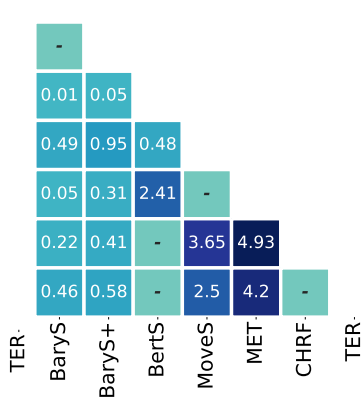

Figure 12: Correctness
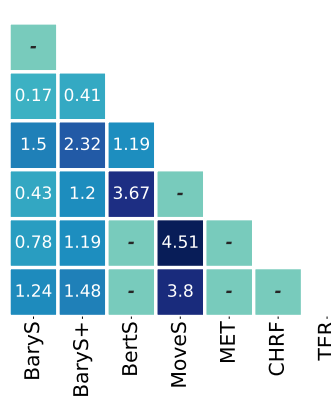
Figure 13: Data Coverage

Figure 14: Relevance

Figure 15: p values of the William significance test on data2text generation.