`InfoLM`: A New Metric to Evaluate Summarization & Data2Text Generation

Pierre Colombo[†*], **Chloé Clavel**[*],**Pablo Piantanida**[†]

[*]Laboratoire des Signaux et Systèmes (L2S), CentraleSupelec CNRS Universite Paris-Saclay
[*]Télécom ParisTech, Université Paris Saclay

`pierre.colombo@centralesupelec.fr`
`chloe.clavel@telecom-paris.fr`
`pablo.piantanida@centralesupelec.fr`

Abstract

Assessing the quality of natural language generation systems through human annotation is very expensive. Additionally, human annotation campaigns are time-consuming and include non-reusable human labour. In practice, researchers rely on automatic metrics as a proxy of quality. In the last decade, many string-based metrics (*e.g.*, `BLEU`) have been introduced. However, such metrics usually rely on exact matches and thus, do not robustly handle synonyms. In this paper, we introduce `InfoLM` a family of untrained metrics that can be viewed as a string-based metric that addresses the aforementioned flaws thanks to a pre-trained masked language model. This family of metrics also makes use of information measures allowing the adaptation of `InfoLM` to various evaluation criteria. Using direct assessment, we demonstrate that `InfoLM` achieves statistically significant improvement and over 10 points of correlation gains in many configurations on both summarization and data2text generation.

1 Introduction

A plethora of applications of natural language processing (NLP) performs text-to-text transformation (Mellish and Dale 1998; Belz and Reiter 2006; Specia, Scarton, and Paetzold 2018). Given an input, these systems are required to produce an output text that is coherent, readable and informative. Due to both high annotation costs and time, researchers tend to rely on automatic evaluation to compare the outputs of such systems. Reference-based automatic evaluation relies on comparing a candidate text produced by the NLG system and one or multiple reference texts ('gold standard') created by a human annotator. Generic automatic evaluation of NLG is a huge challenge as it requires building a metric that evaluates the similarity between a candidate and one or several gold-standard reference texts. However, the definition of success criteria is task-specific: as an example, evaluation of text summarization focuses on content, coherence, grammatically, conciseness, and readability (Mani 2001), whereas machine translation focuses on fidelity, fluency and adequacy of the translation (Hovy 1999) and data2text generation (Gardent et al. 2017) consider criteria such as data coverage, correctness and text structure.

Automatic text evaluation metrics fall into two categories: metrics that are trained to maximise their correlations using human annotation (*e.g.*, `RUSE` (Shimanaka, Kajiwara, and Komachi 2018), `BEER` (Stanojević and Sima'an 2014), `BLEND` (Ma et al. 2017)) and untrained metrics (*e.g.*, `BLEU` (Papineni et al. 2002), `ROUGE` (Lin 2004), `BERTSCORE` (Zhang et al. 2019), `DepthScore` (Staerman et al. 2021b), `BaryScore` (Colombo et al. 2021b), `MOVERSCORE` (Zhao et al. 2019) `Word Mover Distance` (Kusner et al. 2015)). In this work, we focus on untrained metrics as trained metrics may not generalize well to new data (existing labelled corpora are of small size). Two categories of untrained metrics can be distinguished: word or character based-metrics that compute a score based on string representation and embedding-based metrics that rely on a continuous representation. String-based metrics (*e.g.*, `BLEU`) often fail to robustly match paraphrases (Reiter and Belz 2009) as they mainly focus on the surface form as opposed to embedding-based metrics relying on continuous representations.

In this paper, we introduce `InfoLM` a family of new untrained metrics to evaluate text summarization and data2text generation. At the highest level `InfoLM` key components include: (1) a pre-trained masked language model (`PMLM`) that is used to compute two discrete probability distributions over the vocabulary. They represent the probability of observing each token of the vocabulary given the candidate and the reference sentence, respectively. (2) A contrast function $\mathcal{I}$ that is used to measure the dissimilarity between aforementioned probability distributions. `InfoLM` differs from existing BERT-based metrics (*e.g.* `BERTSCORE`, `MOVERSCORE`) as it directly relies on the `PMLM` which outputs discrete probability distributions. Thus `InfoLM` does neither require to arbitrarily select one or several specific layers (*e.g.* `BERTSCORE` relies on the 9th layer for `bert-base-uncased`), nor involves selecting arbitrary aggregations technics (*e.g.* Power Mean (Rücklé et al. 2018) for `MOVERSCORE`). As `InfoLM` relies on statistics on tokens it can also be seen as a string-based metric. However, it does not suffer from common pitfalls of string-based metrics (*e.g.* synonyms, need of an exact string-match) as the `PMLM` also allows ones to assign a high score to paraphrases and to capture distant dependencies.

Contributions Our contributions are summarized below:
(1) *A set of novel metrics to automatically evaluate summarization and data2text generation.* In this work, we introduce `InfoLM` which overcomes the common pitfall of string matching metrics and does not require to select a layer, nor to rely on a arbitrary aggregation function. `InfoLM` combines a pre-trained model and a contrast function denoted by $\mathcal{I}$ between two discrete probability distributions. We explore the use of different choices of contrast functions such as f-divergences, $\mathcal{L}_p$ distances or Fisher-Rao distances. (2) *Tasks.* First, we demonstrate on both summarization and data2text that `InfoLM` is better suited than concurrent metrics. A comparison is conducted, using multiple correlation measures with human judgment both at the text and system level. Second, we dissect `InfoLM` to better understand the relative importance of each component (*e.g.* calibration, sensibility to the change of information measures).

2 Problem statement and related work

In this section, we start by introducing notations and formulate the problem of both evaluating text generation and metrics. Then, we identify and present the most relevant related work and the existing approaches for the studied tasks.

2.1 Problem statement

NLG evaluation. Given a dataset $\mathcal{D} = \{\boldsymbol{x_i}, \{\boldsymbol{y_i^s}, h(\boldsymbol{x_i}, \boldsymbol{y_i^s})\}_{s=1}^S\}_{i=1}^N$ where $\boldsymbol{x_i}$ is the i-th reference text; $\boldsymbol{y_i^s}$ is the i-th candidate text generated by the s-th NLG system; N is the number of texts in the dataset and S the number of systems available. The vector $\boldsymbol{x_i} = (x_1, \cdots, x_M)$ is composed of M tokens (*e.g.*, words or subwords) and $\boldsymbol{y_i^s} = (y_1^s, \cdots, y_L^s)$ is composed of L tokens[1]. The set of tokens (vocabulary) is denoted as Ω, $\mathbf{T}$ denotes the set of possible texts. $h(\boldsymbol{x_i}, \boldsymbol{y_i^s}) \in \mathcal{R}^+$ is the score associated by a human annotator to the candidate text $\boldsymbol{y_i^s}$ when comparing it with the reference text $\boldsymbol{x_i}$. We aim at building an evaluation metric f such that $f(\boldsymbol{x_i}, \boldsymbol{y_i}) \in \mathbb{R}^+$.

Evaluating evaluation metrics. To assess the relevance of an evaluation metric f, correlation with human judgment is considered to be one of the most important criteria (Koehn 2009; Specia, Raj, and Turchi 2010; Chatzikoumi 2020). Debate on the relative merits of different correlations for the evaluation of automatic metrics is ongoing but classical correlation measures are Pearson (Leusch et al. 2003), Spearman (Melamed, Green, and Turian 2003) or Kendall (Kendall 1938) tests. Two meta-evaluation strategies are commonly used: (1) text-level correlation or (2) system-level correlation. Formally, the text-level correlation $C_{t,f}$ is computed as follows:

$$C_{t,f} \triangleq \frac{1}{N} \sum_{i=1}^N K(\mathbf{F}_i^t, \mathbf{H}_i^t), \qquad (1)$$

where $\mathbf{F}_i = \left[f(\boldsymbol{x_i}, \boldsymbol{y_i^1}), \cdots, f(\boldsymbol{x_i}, \boldsymbol{y_i^S}) \right]$ and $\mathbf{H}_i = \left[h(\boldsymbol{x_i}, \boldsymbol{y_i^1}), \cdots, h(\boldsymbol{x_i}, \boldsymbol{y_i^S}) \right]$ are the vectors composed of

---

[1] The reference and candidate text can be composed of several sentences as it is the case in summarization.

---

scores assigned by the automatic metric f and the human respectively. and $K : \mathbb{R}^N \times \mathbb{R}^N \to [-1, 1]$ is the chosen correlation measure (*e.g.*, Pearson, Kendall or Spearman). Similarly, the system level correlation $C_{sy,f}$ is obtained by

$$C_{sy,f} \triangleq K(\mathbf{F}^{sy}, \mathbf{H}^{sy}), \qquad (2)$$

$$\mathbf{F}^{sy} = \left[\frac{1}{N} \sum_{i=1}^N f(\boldsymbol{x_i}, \boldsymbol{y_i^1}), \ldots, \frac{1}{N} \sum_{i=1}^N f(\boldsymbol{x_i}, \boldsymbol{y_i^S}) \right]$$

$$\mathbf{H}^{sy} = \left[\frac{1}{N} \sum_{i=1}^N h(\boldsymbol{x_i}, \boldsymbol{y_i^1}), \ldots, \frac{1}{N} \sum_{i=1}^N h(\boldsymbol{x_i}, \boldsymbol{y_i^S}) \right],$$

where the latter are the vectors composed of the averaged scores assigned by f and the human, respectively. For the significance analysis, we follow Graham and Baldwin (2014) and use a William test to validate a significant improvement for dependent correlations (Steiger 1980).

2.2 Existing metrics

String-based metrics Two types of string-based metrics exist: *N-Grams matching* and *Edit distance-based* metrics. *N-Grams matching metrics* count the number of N-grams in common between the candidate text and the reference text. The three most-used metrics are `BLEU`, `ROUGE` and `METEOR` (Banerjee and Lavie 2005). If no N-gram is in common between the input text candidate and the reference, these metrics fail to produce meaningful scores. The second category of metrics gathers *edit distance-based metrics*. They measure the number of basic operations such as edit/delete/insert to measure semantic equivalence. Variants include `TER` (Snover et al. 2006), `CDER` (Leusch, Ueffing, and Ney 2006), `EED` (Stanchev, Wang, and Ney 2019), `CHARACTER` (Wang et al. 2016). Edit distance-based metrics do not handle synonyms and focus on surface form. `InfoLM` can be seen as string-based but do not suffer from the aforementioned matching problem and can handle synonyms as it relies on a `PMLM`.

Embedding-based metrics Another class of metrics relies on word embeddings. These metrics either use static words embeddings such as word2vec (Mikolov et al. 2013) or contextualized embeddings such as ELMO (Peters et al. 2018), BERT (Devlin et al. 2018) and its variants (Sanh et al. 2019; Liu et al. 2019). Among the most popular metrics, we can mention `MOVERSCORE`, `BERTSCORE`, `WMD` (Kusner et al. 2015) , `WMDO` (Chow, Specia, and Madhyastha 2019). Different from these approaches `InfoLM` relies on a language model and work with discrete probability distributions instead of continuous representations.

Learning-based metrics Various trained metrics have been proposed such as `BEER`, `BEND`, `RUSE`, `CIDER` (Vedantam, Lawrence Zitnick, and Parikh 2015). These methods rely on train/dev/test sets composed of human evaluations. `InfoLM` does not require any training step and relies on a frozen `PMLM`.

`PMLM` as a metric. To the best of our knowledge, using a `PMLM` (*i.e*, without further training) as a reference-based

automatic metric remains overlooked. The closest use we found was to rely on autoregressive models, such as GPT-2 (Radford et al. 2019), to compute the generated sentence perplexity and assess its fluency. Researchers mainly focused on the use of the learnt embedding of the PMLM. However, it remains an open question to find a reliable layer aggregation mechanism (*e.g*, BERTSCORE arbitrary selects a layer based on a chosen dataset, MOVERSCORE uses the 5 last layers). InfoLM addresses this aggregation issue by relying on the PMLM.

2.3 Masked language modelling

Language models based on masked language pre-training objectives Devlin et al. (2018); Liu et al. (2019) aim at reconstructing a corrupt version $[\boldsymbol{x}]$ of an input text $\boldsymbol{x}$ by minimizing a cross-entropy loss. This corrupted context corresponds to a "local view" view of the sentence. To ensure fair comparison, we do not use existing alternatives (*e.g.* GPT-2 based models (Radford et al. 2019)) as concurrent works (Zhang et al. 2019; Zhao et al. 2019) rely on PMLM.

3 InfoLM

In this section, we first introduce a novel family of metrics called InfoLM and then detail the different components of these novel metrics.

Notations We denote by $\theta \in \Theta$, the parameter of the PMLM, $\mathcal{T}$ its temperature and $\mathcal{I} : [0,1]^{|\boldsymbol{\Omega}|} \times [0,1]^{|\boldsymbol{\Omega}|}$ an information measure (see Section 3.3) which quantifies the similarity between two discrete distributions.

3.1 Motivations & Definitions

A PMLM has learnt the empirical distribution of a large text corpus. Given a text $\boldsymbol{x}$, the corrupted context with a mask at position j is denoted $[\boldsymbol{x}]^j$, the LM predicts a distribution $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|[\boldsymbol{x}]^j; \theta; \mathcal{T})$ over the vocabulary $\boldsymbol{\Omega}$ given the masked context. As an example, for a masked input sentence $[\boldsymbol{x}]^1 =$ "The [MASK] was delicious", a pretrained model could place high probabilities on tokens "food", "meal" and low probability on "the". It is worth noting that $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|[\boldsymbol{x}]^1; \theta; \mathcal{T})$ represents the probability of observing each token of the vocabulary given the masked input $[\boldsymbol{x}]^1$.

Definition 3.1 (*Equivalence for masked contexts*). Given $\mathcal{I}$, two masked contexts $[\boldsymbol{x}]^j$, $[\boldsymbol{y}]^k$ from input texts $\boldsymbol{x}$, $\boldsymbol{y}$, with masks at positions j and k respectively, are equivalent (denoted $[\boldsymbol{x}]^j \overset{\mathcal{I}}{\sim} [\boldsymbol{y}]^k$) if the two predicted discrete distributions given by the PMLM, namely $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|[\boldsymbol{x}]^j; \theta; \mathcal{T})$ and $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|[\boldsymbol{y}]^k; \theta; \mathcal{T})$, are similar. Formally, $[\boldsymbol{x}]^j \overset{\mathcal{I}}{\sim} [\boldsymbol{y}]^k$ if $\mathcal{I}\left[p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|[\boldsymbol{x}]^j; \theta; \mathcal{T}), p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|[\boldsymbol{y}]^k; \theta; \mathcal{T})\right] \approx 0$.

Remark. We have the intuition that two similar sentences will share several pairs of equivalent masked contexts. At this point, we make no claim on the relationship between equivalence and the masked context similarity.

In this work, we make the hypothesis that two similar sentences $\boldsymbol{x}, \boldsymbol{y}$ will share multiple equivalent masked contexts. However, pairwise comparisons of all the pairs of individual masked contexts are prohibitively expensive ($\mathcal{O}(L \times M)$

comparisons) when considering long texts. Motivated by efficiency, we instead propose to work with two "global views" of the sentences that are well-formed probability distributions and are obtained through the aggregation of individual PMLM predictions. Aggregated probabilities for $\boldsymbol{x}$ and $\boldsymbol{y}$ are denoted $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x}; \theta; \mathcal{T})$ and $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y}; \theta; \mathcal{T})$ respectively.

Definition 3.2 (*Similarity for texts*). Given $\mathcal{I}$, two texts $\boldsymbol{x}, \boldsymbol{y}$ are said to be similar (denoted $\boldsymbol{x} \overset{\mathcal{I}}{\approx} \boldsymbol{y}$) if $\mathcal{I}\left[p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x}; \theta; \mathcal{T}), p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y}; \theta; \mathcal{T})\right] \approx 0$ where $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x}; \theta; \mathcal{T})$ and $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y}; \theta; \mathcal{T})$ denotes the aggregated individual masked context predictions.

3.2 InfoLM

Overview InfoLM uses the notion of similarity given in Definition 3.2. Given a reference text $\boldsymbol{x_i}$ together with a candidate text $\boldsymbol{y_i^s}$, InfoLM recursively masks each token position of both $\boldsymbol{x_i}$ and $\boldsymbol{y_i^s}$ to obtain individual masked contexts. By relying on a PMLM, InfoLM predicts one distribution for each individual masked contexts. The resulting distributions are then averaged (we refer to this operation "bag of distributions") to obtain $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y_i^s}; \theta; \mathcal{T})$ and $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x_i}; \theta; \mathcal{T})$. The final step involves comparing two well formed discrete probability distributions $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y_i^s}; \theta; \mathcal{T})$ and $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x_i}; \theta; \mathcal{T})$ through a measure of information $\mathcal{I}$. InfoLM writes as:

$$\text{InfoLM}(\boldsymbol{x_i}, \boldsymbol{y_i^s}) \triangleq \mathcal{I}\left[p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x_i}; \theta; \mathcal{T}), p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y_i^s}; \theta; \mathcal{T})\right]. \quad (3)$$

Remark. It is worth to emphasize that $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x_i}; \theta; \mathcal{T})$ and $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y_i^s}; \theta; \mathcal{T})$ are two well formed discrete probability distributions. They represent the probability of observing each token of the vocabulary given the candidate and the reference sentence, respectively.

Aggregation Procedure Rare tokens can be more indicative of text similarity than common tokens (Banerjee and Lavie 2005). Thus, for the aggregation of the individual masked contexts, we propose to compute a weighted "bag of distributions" where the weights are normalized measures of the importance of each token. In practice, $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x}; \theta; \mathcal{T})$ and $p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y_i^s}; \theta; \mathcal{T})$ write as:

$$p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x}; \theta; \mathcal{T}) \triangleq \sum_{k=1}^{M} \gamma_k \times p_\theta(\cdot|[\boldsymbol{x}]^k; \theta; \mathcal{T}),$$

$$p_{\boldsymbol{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y_i^s}; \theta; \mathcal{T}) \triangleq \sum_{k=1}^{N} \tilde{\gamma}_k \times p_\theta(\cdot|[\boldsymbol{y_i^s}]^k; \theta; \mathcal{T}),$$

where $\tilde{\gamma}_k$ and γ_k are measures of the importance of the k-th token in the candidate and reference text, respectively, i.e., satisfying $\sum_{j=1}^{M} \gamma_j = \sum_{j=1}^{N} \tilde{\gamma}_j = 1$. These are computed using the inverse document frequency scores determined at the corpus level (Zhao et al. 2019; Kusner et al. 2015).

LM Calibration. Modern deep neural networks are overconfident (Guo et al. 2017). To re-calibrate language models several techniques have been proposed (*e.g* entropy rates, (Braverman et al. 2020) temperature scaling (Platt et al. 1999), contextual calibration (Zhao et al. 2021)). Here, we

choose to study how calibration affects `InfoLM` by relying on temperature[2] scaling motivated by simplicity and speed.

3.3 Information measures

In this work, we focus on comparing a pair of discrete probability distributions through information measures (see Basseville (2013) for an exhaustive study). We rely on two types of information measures: *divergences* and *distances*. The divergence is a measure of dissimilarity that is always positive or equal to zero if (and only if) the two considered distributions are strictly identical. We call distance, a function that is symmetric, positive, respects the triangle inequality and is equal to zero if (and only if) the two considered distributions are strictly identical. We will use information measures that belong to either Csiszar f-divergences or that are distances.

Divergence measures   Various divergence measures have been proposed for a large variety of applications (Basseville 2013). The full expression of the studied divergences can be found in Table 1. We focus here on three families of divergences α Divergences, γ Divergences and AB Divergences. Note that there exist other families of divergences such as Bregman divergence (Bregman 1967), β divergences (Basu et al. 1998), Chernoff divergence (Chernoff et al. 1952).

α-**Divergences.** This divergence was introduced by Rényi et al. (1961) and are a special case of the f-divergences (Csiszár 1967). They are widely used in variational inference (Li and Turner 2016) and closely related to Rényi divergences but are not a special case. From Table 1 we note special cases of α-Divergences: (i) Kullback-Leiber (KL) is recovered by letting $\alpha \to 1$, (ii) Hellinger distance (Hellinger 1909) follows by choosing $\alpha = 0.5$. For this family, α weights the influence of $\frac{p}{q}$.

γ-**Divergences.** This divergence has been introduced by (Fujisawa and Eguchi 2008) as a scale-invariant modification of the robust β-divergences.[3] For the γ divergences the parameter β is used to control the importance of the element of small probabilities (*e.g.*, outliers in some scenarios, tokens with low probability in our case). If $\beta > 1$, the importance of large q_i is reduced which gives more weights to the outliers. Special cases include the L_2 distance (i.e., $\beta = 2$) and KL divergence (i.e., $\beta \to 1$).

AB-**Divergences.** The family of AB-divergences is flexible and allows to respectively control the mass coverage or the robustness. Cichocki, Cruces, and Amari (2011) propose to use AB divergences. As can be seen in Table 1 these divergences have two parameters α, β. It allows to tune the mass coverage and the robustness independently. The β-divergence is obtained by choosing $\alpha = 1, \beta \in \mathbb{R}$.

From information divergences to discrimination. For our application, we would like to produce a metric between two texts regardless of the source (system or human). Thus we are interested in symmetric divergence: such divergences are called discrimination. To obtain discrimination

two tricks are commonly applied either the Jeffrey's symmetrization, which averages $\text{KL}(p\|q)$ and $\text{KL}(q\|p)$), or the Jensen's symmetrization, which averages $\text{KL}(p\|\frac{p+q}{2})$ and $\text{KL}(q\|\frac{p+q}{2})$. We choose to use Jeffreys symmetrization as it does not require computing $\frac{p+q}{2}$. The symmetric KL with Jeffrey's symmetrization is denoted JS.

Distances $\mathcal{L}_p$ **distances.** The $\mathcal{L}_p$ distances $p \in \mathbb{R}_+$ can be used to measure the similarity between two distributions and we restrict ourselves to $p \in \{1, 2, +\infty\}$.
Fisher-Rao distance. The Fisher-Rao distance (R) represents the Geodesic Distance (Rao 1987) between two distributions. Interestingly, this distance remains overlooked in the ML community but has been recently used to achieve robustness against adversarial attacks (Picot et al. 2021).

Connection with string matching metrics.   We adopt the following notations $p_{\mathbf{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x_i}; \theta; \mathcal{T}) = [p_0, \cdots, p_{|\Omega|}]$ and $p_{\mathbf{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y_i^s}; \theta; \mathcal{T}) = [q_0, \cdots, q_{|\Omega|}]$. First, let us consider two texts $\boldsymbol{x_i}, \boldsymbol{y_i^s}$ such that $\boldsymbol{x_i} \overset{\mathcal{I}}{\approx} \boldsymbol{y_i^s}$. `InfoLM` with $\mathcal{L}_\infty$ is closed to 0 if $\forall i \in [1, |\Omega|]$ $p_i \approx q_i$. It means that all likely tokens (according to the `PMLM`) when considering $\boldsymbol{x_i}$ are also likely when considering $\boldsymbol{y_i^s}$. For string matching metrics, it corresponds to a *perfect match* between $\boldsymbol{x_i}$ and $\boldsymbol{y_i^s}$. Second, let us consider $\boldsymbol{x_i}, \boldsymbol{y_i^s}$ such that $\mathcal{I}\left[p_{\mathbf{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{x_i}; \theta; \mathcal{T}), p_{\mathbf{\Omega}|\mathbf{T}}(\cdot|\boldsymbol{y_i^s}; \theta; \mathcal{T})\right] \approx 1$ (dissimilar texts) and a measure of information that relies on product of p_i and q_i (*e.g* Fisher-Rao). In this case, $\forall i \in [1, |\Omega|]$ $p_i \times q_i \approx 0$ thus all likely tokens when considering $\boldsymbol{x_i}$ are unlikely when considering $\boldsymbol{y_i^s}$ (the converse it true as well). For string matching metrics this corresponds to *no match* among the sub-strings of $\boldsymbol{x_i}$ and $\boldsymbol{y_i^s}$.

4 Experimental Frameworks

In this section, we describe our experimental setting. We present the tasks and the baselines metrics use for each task.

4.1 Text summarization

Text summarization aims at compressing long texts into fluent, short sentences that preserve the salient information.
Datasets. To compare the different metrics previous work (Bhandari et al. 2020; Fabbri et al. 2020) either relies on the TAC datasets (Dang and Owczarzak 2008; McNamee and Dang 2009) or on new summarization datasets extracted from CNN/DailyMail (Nallapati et al. 2016). As pointed out by Peyrard (2019); Bhandari et al. (2020), TAC datasets are old and contain flaws (*e.g* systems used to generate summaries were of poor quality), we choose to work with the newly assemble dataset from CNN/Daily News proposed in Bhandari et al. (2020). This dataset gathers 11,490 summaries and annotations are carried using the pyramid method (Nenkova and Passonneau 2004).
Metrics. For text summarization, perhaps the most known metrics are `ROUGE` and its extensions (Ng and Abrecht 2015), or `METEOR` and its variants (Denkowski and Lavie 2014). Recently, a new set of metrics (*e.g* `BERTSCORE`, `MOVERSCORE`) have been applied to text summarization.

---

[2]When $\mathcal{T} \to 0$ one token receive all the probability mass, and when $\mathcal{T} \to \infty$, the probability becomes uniform.

[3]In our setting we work we normalised distributions, thus scale invariance is not a mandatory property. It is worth mentioning as it could cause practical issues when optimising our metric.

Name	Notation	Domain	Expression
α-divergence (Csiszár 1967)	$\mathcal{D}_\alpha$	$\alpha \notin \{0,1\}$	$\frac{1}{\alpha(\alpha-1)}(1 - \sum q_i^{1-\alpha} p_i^\alpha)$
γ divergence (Fujisawa and Eguchi 2008)	$\mathcal{D}_\gamma^\beta$	$\beta \notin \{0,-1\}$	$\frac{1}{\beta(\beta+1)} \log \sum p_i^{\beta+1} + \frac{1}{\beta+1} \log \sum q_i^{\beta+1} - \frac{1}{\beta} \log \sum p_i q_i^\beta$
AB Divergence (Cichocki, Cruces, and Amari 2011)	$\mathcal{D}_{sAB}^{\alpha,\beta}$	$(\alpha,\beta) \in (\mathbb{R}^*)^2$ $\beta + \alpha \neq 0$	$\frac{1}{\beta(\beta+\alpha)} \log \sum p_i^{\beta+\alpha} + \frac{1}{\beta+\alpha} \log \sum q_i^{\beta+\alpha} - \frac{1}{\beta} \log \sum p_i^\alpha q_i^\beta$
$\mathcal{L}_1$ distance	$\mathcal{L}_1$		$\sum |p_i - q_i|$
$\mathcal{L}_2$ distance	$\mathcal{L}_2$		$\sqrt{\sum (p_i - q_i)^2}$
$\mathcal{L}_\infty$ distance	$\mathcal{L}_\infty$		$\max_i |p_i - q_i|$
Fisher-Rao distance	R		$\frac{2}{\pi} \arccos \sum \sqrt{p_i \times q_i}$

Table 1: Expression of the divergences (upper group) and distance between two positives measures $\mathbf{p} = (p_1, \cdots, p_N)$ and $\mathbf{q} = (q_1, \cdots, q_N)$ as well as the definition domain (see (Regli and Silva 2018)). We omit the index in the summations.

4.2 Data2Text generation

Prior works mainly rely on two task-oriented dialogue datasets (*i.e.*, BAGEL (Mairesse et al. 2010), SFHOTEL (Wen et al. 2015)). As sentence generated in these data-sets are unlikely to be representative of the progress of recent NLG systems we instead rely on a different dataset coming from the WebNLG2020 challenge (Gardent et al. 2017). Given the following example of triple: *(John_Blaha birth-Date 1942_08_26) (John_Blaha birthPlace San_Antonio) (John_E_Blaha job Pilot)* the goal is to generate *John Blaha, born in San Antonio on 1942-08-26, worked as a pilot.*

Annotations. The WebNLG task is evaluated by human annotators along four different axes: (1) Data Coverage: Are all the descriptions presented in the data included in the text? (2) Relevance: Does the text contains only predicates found in the data? (3) Correctness: Are predicates found in the data correctly mentioned and adequately introduced? (4) Text structure: Is the produced text well-structured, grammatically correct and written in acceptable English?, (5) Fluency: Does the text progress naturally? Is it easy to understand? Is it a coherent whole?

Metrics. For this task, organisers rely on untrained metrics (*e.g.* BLEU, METEOR, TER, BERTSCORE) to compare the performance of the candidate systems. Thus, we will focus on system-level correlation.

5 Numerical Results

In this section, we study the performance of InfoLM on both text summarization and data2text generation.

5.1 Results on Text Summarization

General Analysis. Figure 8 gathers the results of the correlation study between scores produced by different metrics and human judgement (*i.e.* pyramid score). We can reproduce results from Bhandari et al. (2020). We observe a different behavior depending on the type of systems to be evaluated (*e.g.*, abstractive or extractive) and the chosen correlation coefficient. We observe that InfoLM with $\mathcal{D}_{AB}$ or with $\mathcal{R}$ outperforms other BERT-based metrics such as MOVERSCORE or BERTSCORE (*e.g.*, it is worth noting that both metrics perform poorly at the text or system-level when considering outputs from extractive systems). $\mathcal{D}_{AB}$ largely

outperforms n-gram matching metrics (*e.g.*, ROUGE metrics) on all datasets when measuring correlation with the Kendall τ and in almost all configurations (except when considering abstractive outputs at the system level) when using the Pearson r. It is worth noting the overall good performance of the parameter-free Fisher-Rao distance.

Choice of information geometric measure for InfoLM. In Figure 8, we can observe two different types of groups depending on the global behaviour. First we notice that using $\mathcal{L}_p$, $p \in \{1, 2, \ldots, +\infty\}$ leads to poor performances in many configurations. Good performance of $\mathcal{L}_\infty$ in some configurations is surprising as $\mathcal{L}_\infty$ is extremely selective (*i.e.* $\mathcal{L}_\infty$ computes $\max_i |p_i - q_i|$). As output produced by the PMLM is sparse, $\max_i |p_i - q_i|$ correspond to one likely word in one sentence and not likely at all in the other. The second group gathers JS, $\mathcal{R}$, $\mathcal{D}_\alpha$, $\mathcal{D}_\beta$ and $\mathcal{D}_{AB}$ and achieves the best performance overall. $\mathcal{D}_\alpha$ and $\mathcal{D}_{AB}$ achieve similar performance suggesting that the flexibility (*e.g.*, robustness to outliers) introduced by the β parameter in $\mathcal{D}_{AB}$ is not useful in our task. This observation is strengthened by the lower performance of $\mathcal{D}_\beta$. The difference of results between the two measures is due to the flexibility introduced by α (*i.e.*, α controls the relative importance of the ration $\frac{p_i}{q_i}$) which can be interpreted in our case as the ability to control the importance attributed to less likely words (Jalalzai et al.).

Takeaways. The best performing metric is obtained with $\mathcal{D}_{AB}$. The Fisher-Rao distance, denoted by $\mathcal{R}$, achieves good performance in many scenarios and has the advantage to be parameter-free.

5.2 Results on Data2Text

Global Analysis: Table 3 gathers results of the correlation analysis of the metrics with human judgements following the five different axes. We observe that the five considered criteria of annotations are not independent: text structure and fluency achieve a strong correlation coefficient (> 98). Additionally, all metrics achieve similar results when the correlation is computed on these two criteria. We observe that the best performing group of metric is based on InfoLM followed by metrics based on continuous representation from BERT (*i.e.*, MOVERSCORE and BERTSCORE) followed by N-gram matching metrics. Regarding correct-
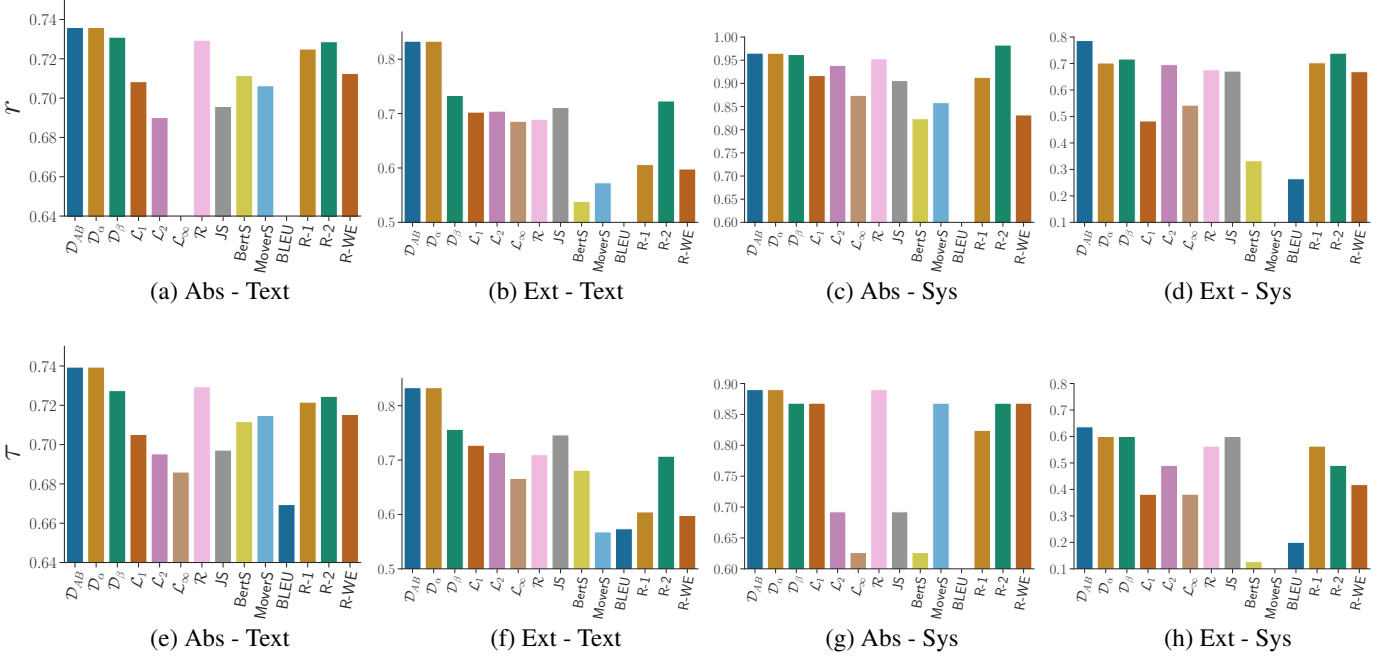
(a) Abs - Text (b) Ext - Text (c) Abs - Sys (d) Ext - Sys

(e) Abs - Text (f) Ext - Text (g) Abs - Sys (h) Ext - Sys

Figure 1: Results of the correlation between metrics and human judgments on the CNN dataset. First raw reports correlation as measured by the Person (r) and second raw focus on Kendall (τ) coefficient. In this experiment, parameter are optimized for each criterion. The results on with Spearman (ρ) can be found in Appendix B.5.

ness, data coverage and relevance, we observe that both $\mathcal{D}_{AB}$ and $\mathcal{D}_\alpha$ achieve the best results on almost all correlation coefficients. On data coverage, `InfoLM` achieves improvement up to 17 points in correlation compared to both BERT based or N-gram matching metrics. Regarding fluency and text structure, Fisher-Rao distance works better and slightly outperforms the second-best performing metric, namely `BERTSCORE`.

Takeaways. Similar to summarisation, we observe very low correlation for L_p, $p \in \{1, 2, \ldots, +\infty\}$. We also observe that β-divergences achieve lower results than both α and AB divergences suggesting that, as noticed for summarisation, robustness to unlikely words (*i.e.*, outliers) is less relevant for our task.

5.3 Further Analysis

Correlation between metrics In this experiment, we complete our global analysis by comparing the scores obtained by the different metrics with each other. We want to gain an understanding of how different our metric is from other metrics and how the choice of information geometric measures affects the predictions. Figure 2 gathers the results of the experiment. We observe a high correlation ($r > 88$) between $\mathcal{D}_\alpha$, $\mathcal{D}_\beta$, $\mathcal{D}_{AB}$ and $\mathcal{R}^4$. Interestingly, we observe a lower correlation ($r \approx 70$) with `BERTSCORE` and N-gram matching metrics, e.g., `ROOGE`) whereas `BERTSCORE` achieves a stronger correlation with `ROUGE` ($r \approx 80$).

Takeaways. Through the correlation analysis in Figure 2, we observe the impact of different geometry on `InfoLM` predictions. The correlation analysis shows that the prediction of `InfoLM` when using $\mathcal{D}_\alpha$, $\mathcal{D}_\beta$, $\mathcal{D}_{AB}$ and $\mathcal{R}$ are highly

---

[4]Note that these metrics consider the product of p_i and q_i.

correlated and as illustrated by previous experience achieve high correlation scores which we believe validate our approach. It is worth noting that $\mathcal{R}$ requires no tuning as it is parameter free.
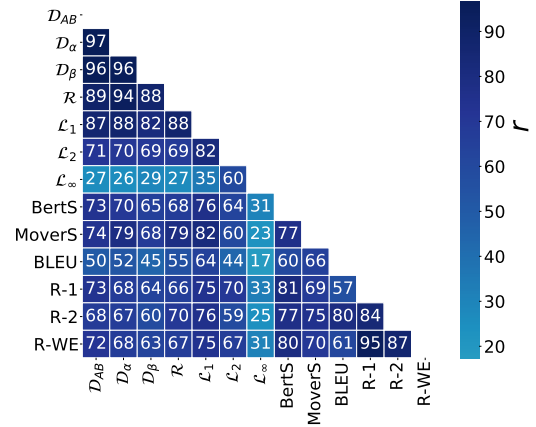

Figure 2: Pearson correlation at the system level between metrics when considering abstractive system outputs.

Score Distributions In Figure 3, we study the text score distribution of different metrics on abstractive summary. The ideal metric would mimic the human score distribution (*i.e. Pyr.*) and be able to distinguish between good and bad quality summaries. The results show that `ROUGE` and `BLEU` struggle to distinguish between between good quality ($Pyr. \geq 0.5$) low quality summaries ($Pyr. \leq 0.5$) which has been reported in Peyrard (2019). We observe that $\mathcal{D}_{AB}$,

Metric	Correctness			Data Coverage			Fluency			Relevance			Text Structure		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
Correct	100.0	100.0	100.0	97.6	85.2	73.3	80.0	81.1	61.6	99.1	89.7	75.0	80.1	80.8	60.0
DataC	85.2	97.6	73.3	100.0	100.0	100.0	71.8	51.7	38.3	96.0	93.8	81.6	71.6	51.4	36.6
Fluency	81.1	80.0	61.6	71.8	51.7	38.3	100.0	100.0	100.0	77.0	61.4	46.6	99.5	99.7	98.3
Relev	89.7	99.1	75.0	96.0	93.8	81.6	77.0	61.4	46.6	100.0	100.0	100.0	77.2	61.1	45.0
TextS	80.8	80.1	60.0	71.6	51.4	36.6	99.5	99.7	98.3	77.2	61.1	45.0	100.0	100.0	100.0
$\mathcal{D}_{AB}$	88.8	**89.3**	**76.6**	**81.8**	**82.6**	**70.0**	86.6	92.0	76.6	**89.8**	**87.9**	73.3	86.6	91.4	75.0
$\mathcal{D}_{\alpha}$	88.8	**89.3**	**76.6**	**81.8**	**82.6**	**70.0**	86.6	92.0	76.6	**89.8**	**87.9**	73.3	86.6	91.4	75.0
$\mathcal{D}_{\beta}$	81.4	50.0	71.6	48.4	79.7	65.0	44.8	84.7	76.6	49.3	72.3	60.0	48.0	83.8	75.0
$\mathcal{L}_1$	75.2	33.8	61.6	32.4	53.8	40.0	22.7	83.5	73.3	32.2	57.9	45.0	25.6	83.2	71.6
$\mathcal{R}$	**89.7**	86.0	75.0	78.7	70.5	51.6	**93.3**	**95.7**	**85.3**	87.6	84.4	70.0	**92.4**	**93.8**	**81.6**
JS	79.4	81.1	70.0	69.3	75.5	60.0	89.4	91.4	75.0	81.7	70.5	60.0	91.9	91.1	73.3
BertS	85.5	83.4	73.3	74.7	68.2	53.3	92.3	95.5	85.0	83.3	79.4	65.0	91.9	**95.0**	**83.3**
MoverS	84.1	84.1	73.3	78.7	66.2	53.3	91.2	92.1	78.3	82.1	77.4	65.0	90.1	91.4	76.3
BLEU	77.6	66.3	60.0	55.7	50.2	36.6	89.4	90.5	78.3	63.0	65.2	51.6	88.5	89.1	76.6
R-1	80.6	65.0	65.0	61.1	59.6	48.3	76.5	76.3	60.3	64.3	69.2	56.7	75.9	77.5	58.3
METEOR	86.5	66.3	70.0	77.3	50.2	46.6	86.7	90.5	78.3	82.1	65.2	58.6	86.2	89.1	76.6
TER	79.6	78.3	58.0	69.7	58.2	38.0	89.1	93.5	80.0	75.0	70.2	**77.6**	89.5	91.1	78.6

Table 2: Correlation at the system level with human judgement along five different axis: correctness, data coverage, fluency, relevance and text structure for the WebNLG task. Best results by group are underlined, overall best results are bolted.

$\mathcal{R}$ and JS metrics are able to make the distinction. Interestingly, as p increases and the $\mathcal{L}_p$ distances become more selective (*i.e.* focus on one word solely), the $\mathcal{L}_p$ distances struggle to distinguish low from high scoring summaries.

Takeaways. `InfoLM` when combined with $\mathcal{D}_{AB}$, $\mathcal{R}$ and JS is able to distinguish high-scoring from low scoring summaries.
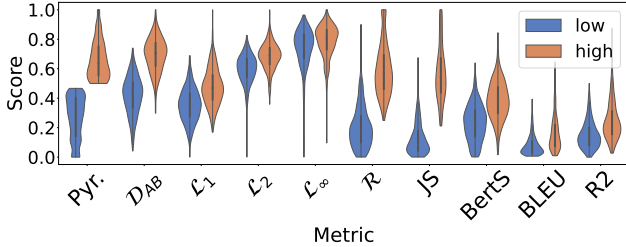
Figure 3: Score distribution of text score when considering abstractive system outputs. $Pyr.$ stands for pyramide score.

Temperature Calibration To study the impact of calibration, we choose to work on system-level correlation and report in Figure 4 the achieved the correlation measured by the different coefficients. We limit our study to the Fisher-Rao distance as it is a parameter-free metric and is among the best-performing metrics of `InfoLM`. Due to space constraints, we report the result on extractive systems only (extractive systems can be found in Figure 5).

Takeaways. Fisher-Rao only considers product $p_i \times q_i$ thus as $\mathcal{T}$ increases and the predicted probability of the `PMLM` becomes more uniform more words are considered and the aggregated distributions become richer in terms of considered words. It is worth noting that when changing the temperature we observe a smooth change in correlation and observe an optimal temperature T which is reached for $T \in [1, 2]$.

It suggests that `InfoLM` benefits from a `PMLM` that is not too selective (case $T \ll 1$). For a specific application, the temperature of `InfoLM` can be tuned to improve correlation and `InfoLM` will likely benefit from well-calibrated `PMLM`.
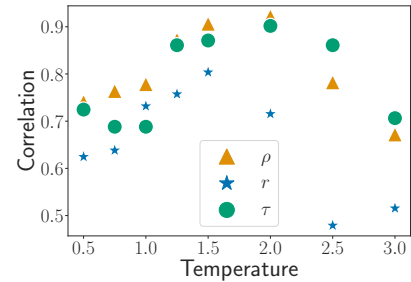
Figure 4: Impact of Calibration on system-level correlation for summarization.

6 Summary and Concluding Remarks

In this work, we presented `InfoLM` that does not require training and it is among the first metrics computing the similarity between two discrete probability distributions over the vocabulary (which is similar to string-based metrics) but also leverages the recent advance in language modeling thanks to a `PMLM`. Our experiments on both summarization and data2text generation demonstrate the validity of our approach. Among available contrast measures, the Fisher-Rao distance is parameter-free and thus, it is easy to use in practice while the AB-Divergence achieves better results but requires to select α and β. Future work includes extending our metrics to new tasks such as translation, image captioning and paraphrase detection, as well as investigating others information geometric measures (*e.g.*, Wasserstein distance (Staerman et al. 2021a), or other types of divergences and Mutual Information (Colombo, Clavel, and Piantanida 2021)) to obtain better correlation with different attributes.

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of ACL workshop*.

Basseville, M. 2013. Divergence measures for statistical data processing—An annotated bibliography. *Signal Processing*, 93(4): 621–633.

Basu, A.; Harris, I. R.; Hjort, N. L.; and Jones, M. 1998. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3): 549–559.

Belz, A.; and Reiter, E. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the EACL*.

Bhandari, M.; Gour, P.; Ashfaq, A.; Liu, P.; and Neubig, G. 2020. Re-evaluating evaluation in text summarization.

Braverman, M.; Chen, X.; Kakade, S.; Narasimhan, K.; Zhang, C.; and Zhang, Y. 2020. Calibration, Entropy Rates, and Memory in Language Models. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th ICML*, volume 119 of *PMLR*, 1089–1099. PMLR.

Bregman, L. M. 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3).

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

Chernoff, H.; et al. 1952. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4).

Chow, J.; Specia, L.; and Madhyastha, P. 2019. WMDO: Fluency-based Word Mover's Distance for Machine Translation Evaluation. In *Proceedings of the Fourth WMT (Volume 2: Shared Task Papers, Day 1)*, 494–500. Florence, Italy: ACL.

Cichocki, A.; Cruces, S.; and Amari, S.-i. 2011. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1): 134–170.

Colombo, P.; Clavel, C.; and Piantanida, P. 2021. A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations.

Colombo, P.; Staerman, G.; Clavel, C.; and Piantanida, P. 2021. Automatic text evaluation through the lens of wasserstein barycenters. *arXiv preprint arXiv:2108.12463*.

Csiszár, I. 1967. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2: 229–318.

Dang, H. T.; and Owczarzak, K. 2008. Overview of the TAC 2008 Update Summarization Task. In *TAC*.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fabbri, A. R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; and Radev, D. 2020. Summeval: Re-evaluating summarization evaluation.

Fujisawa, H.; and Eguchi, S. 2008. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9): 2053–2081.

Gardent, C.; Shimorina, A.; Narayan, S.; and Perez-Beltrachini, L. 2017. Creating training corpora for nlg micro-planning. In *55th ACL (ACL)*.

Graham, Y.; and Baldwin, T. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 EMNLP (EMNLP)*, 172–176.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *ICML*.

Hellinger, E. 1909. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136).

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

Jalalzai, H.; Colombo, P.; Clavel, C.; Gaussier, É.; Varni, G.; Vignon, E.; and Sabourin, A. ???? Heavy-tailed representations, text polarity classification & data augmentation.

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

Koehn, P. 2009. *Statistical machine translation*. Cambridge University Press.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *ICML*.

Leusch, G.; Ueffing, N.; and Ney, H. 2006. CDER: Efficient MT Evaluation Using Block Movements. In *11th Conference of the EACL*. Trento, Italy: ACL.

Leusch, G.; Ueffing, N.; Ney, H.; et al. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of Mt Summit IX*, 240–247.

Li, Y.; and Turner, R. E. 2016. R\'enyi Divergence Variational Inference. *arXiv preprint arXiv:1602.02311*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: ACL.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ma, Q.; Graham, Y.; Wang, S.; and Liu, Q. 2017. Blend: a novel combined MT metric based on direct assessment—casict-dcu submission to WMT17 metrics task. In *Proceedings of the second WMT*, 598–603.

Mairesse, F.; Gasic, M.; Jurcicek, F.; Keizer, S.; Thomson, B.; Yu, K.; and Young, S. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th ACL*, 1552–1561.

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

McNamee, P.; and Dang, H. T. 2009. Overview of the TAC 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, volume 17, 111–113.

Melamed, I. D.; Green, R.; and Turian, J. 2003. Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, 61–63.

Mellish, C.; and Dale, R. 1998. Evaluation in the context of natural language generation. *Computer Speech & Language*, 12(4): 349–373.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Nallapati, R.; Zhou, B.; Gulcehre, C.; Xiang, B.; et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Nenkova, A.; and Passonneau, R. J. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NACCL 2004*, 145–152.

Ng, J.-P.; and Abrecht, V. 2015. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th ACL*, 311–318. Philadelphia, Pennsylvania, USA: ACL.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Peyrard, M. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th ACL*, 5093–5100.

Picot, M.; Messina, F.; Boudiaf, M.; Labeau, F.; Ben Ayed, I.; and Piantanida, P. 2021. Adversarial Robustness via Fisher-Rao Regularization. *arXiv preprint arXiv:*.

Platt, J.; et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3): 61–74.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8).

Rao, C. R. 1987. Differential metrics in probability spaces. *Differential geometry in statistical inference*, 10: 217–240.

Regli, J.-B.; and Silva, R. 2018. Alpha-beta divergence for variational inference. *arXiv preprint arXiv:1805.01045*.

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

Rényi, A.; et al. 1961. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Rücklé, A.; Eger, S.; Peyrard, M.; and Gurevych, I. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shimanaka, H.; Kajiwara, T.; and Komachi, M. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third WMT: Shared Task Papers*, 751–758.

Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; and Makhoul, J. 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.

Specia, L.; Raj, D.; and Turchi, M. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1): 39–50.

Specia, L.; Scarton, C.; and Paetzold, G. H. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1): 1–162.

Staerman, G.; Laforgue, P.; Mozharovskyi, P.; and d'Alché Buc, F. 2021a. When ot meets mom: Robust estimation of wasserstein distance. In *International Conference on Artificial Intelligence and Statistics*, 136–144. PMLR.

Staerman, G.; Mozharovskyi, P.; Clémençon, S.; and d'Alché Buc, F. 2021b. Depth-based pseudo-metrics between probability distributions.

Stanchev, P.; Wang, W.; and Ney, H. 2019. EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth WMT (Volume 2: Shared Task Papers, Day 1)*, 514–520.

Stanojević, M.; and Sima'an, K. 2014. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 414–419. Baltimore, Maryland, USA: ACL.

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

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

Wang, W.; Peter, J.-T.; Rosendahl, H.; and Ney, H. 2016. Character: Translation edit rate on character level. In *Proceedings of the First WMT: Volume 2, Shared Task Papers*, 505–510.

Wen, T.-H.; Gasic, M.; Mrksic, N.; Su, P.-H.; Vandyke, D.; and Young, S. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert.

Zhao, T. Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models.

Zhao, W.; Peyrard, M.; Liu, F.; Gao, Y.; Meyer, C. M.; and Eger, S. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance.