Measuring the Measuring Tools: An Automatic Evaluation of Semantic Metrics for Text Corpora

George Kour*, **Samuel Ackerman***
Orna Raz, Eitan Farchi, Boaz Carmeli, Ateret Anaby-Tavor
IBM Research AI
{gkour, samuel.ackerman}@ibm.com
{ornar, farchi, boazc, atereta}@il.ibm.com

Abstract

The ability to compare the semantic similarity between text corpora is important in a variety of natural language processing applications. However, standard methods for evaluating these metrics have yet to be established. We propose a set of automatic and interpretable measures for assessing the characteristics of corpus-level semantic similarity metrics, allowing sensible comparison of their behavior. We demonstrate the effectiveness of our evaluation measures in capturing fundamental characteristics by evaluating them on a collection of classical and state-of-the-art metrics. Our measures revealed that recently-developed metrics are becoming better in identifying semantic distributional mismatch while classical metrics are more sensitive to perturbations in the surface text levels.

1 Introduction

While there has been a long-standing interest in developing semantic similarity metrics[1] (Rayson and Garside, 2000), measuring how close two text corpora are remains an open problem (Pillutla et al., 2021). Specifically, the recent advances in generative language models have led to an increased interest in the study of content similarity between human and generated language, as a mean for comparing the quality of generative models (Mille et al., 2021; Gehrmann et al., 2022).

The goal of a text corpus' dissimilarity or distance metric is to provide a broad representation of distance across specific linguistic aspects, such as lexical, morphological, syntactic, and semantic (Kilgarriff, 2001). Such metrics are essential for measuring how well corpus-based linguistic analysis generalizes from one data-set to another. This work focuses on semantic similarity metrics.

While one can reasonably measure the semantic distance between two individual sentences (e.g., by calculating the cosine distance between the sentence embeddings), measuring the dissimilarity between two text corpora remains a challenge (Naeem et al., 2020). Corpus-level metrics seek to assess semantic similarity at the group level. for instance, assessing generated text fidelity, diversity, and coverage compared to the reference corpus (Sajjadi et al., 2018). Thus, one common approach for measuring the semantic dissimilarity between two corpora is to compare the densities of their sentences in the embedding space (Pillutla et al., 2021).

However, there are no standard automatic procedures for evaluating the precision and robustness of such similarity metrics. The semi-manual standard approach is to correlate the results of these metrics for human judgement. However, leveraging manual human judgements to construct numeric metrics has significant weaknesses. As we explain in Section 2, human judgements are expensive to obtain, are difficult to aggregate consistently from individual text instances into a corpus-level metric, and can be subjective and non-robust.

To mitigate the dependence on human judgement, controllable synthetic distributions have been used in recent work to evaluate the metric quality (Naeem et al., 2020). For instance, this was done by calculating the distance of synthetic high dimensional samples generated by sampling from Gaussian or mixture-of-Gaussian distributions representing the reference P and target Q data, and then calculating the distance measure by shifting away the two distributions. In this paper, we adopt a middle ground between validating the metric against human judgement on real data and evaluating the metric with synthetic distributions by building "controllable-distance real data corpora" (Section 3). By precisely controlling the content of test corpora, we devised a unified evaluation of desired metric characteristics on real data. This tech-

* denotes equal contribution.

[1] In the context of this paper, a metric is a measure of difference (distance) in the general sense, and may not necessarily satisfy the properties of a metric in mathematical terms.

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 405 - 416

nique allows aggregation of many small-difference judgements that should correspond to what a human would logically decide, to evaluate the distance metric overall in terms of desirable properties. The middle ground thus attempts to reflect human logical judgement in an inexpensive way, while avoiding some of the weaknesses described, such as lack of consistency.

To summarize, our contributions are as follows. First, we present a text similarity evaluation measures that allows researchers to compare the robustness of their newly proposed metrics against existing metrics using the same standards. Second, we evaluate classical and state-of-the-art similarity metrics and show that our benchmark performs well in capturing their known properties. Finally, we provide a pip-installable Python package to compute an extensive set of text dissimilarity metrics, using a unified and simple API[2].

2 Literature Review

The most widely-used method to compute the quality of text similarity metrics investigates the correlation between the scores given by the metric and human judgements. However, human judgement, even on the sentence level, has several shortcomings, mainly that it is expensive and can be inconsistent and subjective (Popescu-Belis, 2003; Lin and Och, 2004; Graham et al., 2017). Also, superficial aspects of the sentences, such as text length or syllables per sentence, may influence human judgements of the semantic similarity (Novikova et al., 2017). Furthermore, though humans may be able judge the relative similarity of a pair of *sentences*, they are usually limited in their ability to make large-scale assessments of a similar type when comparing two *corpora* (i.e., two distributions of sentences) consistently and reliably.

In an attempt to standardize metric evaluation, several competitions and standard datasets containing compared data and human assessment were created for specific tasks, such as translation (Guo et al., 2018; Mathur et al., 2020). However, there is currently a lack of benchmarks against which to assess the semantic similarity between corpora.

Text similarity metrics can be thought of as belonging to several broad and overlapping classes (see e.g., Wang and Dong 2020), which partially depend on the form of the text representation (e.g., token-based or vector embedding). Here, we investigate metrics from three of these classes, comparing corpora based on these aspects: *lexicographical* (statistical properties of words and tokens), *distribution* (densities of sentences represented in the embedding space), and *discriminatability* (ability to classify sentences as belonging to one corpus or the other). The metrics we use are summarized in Table 1.

Lexicographical Statistics These methods have been developed to compare various distributional properties of target text Q with respect to the reference samples P, based on some statistic measures $T(P)$ and $T(Q)$, operating on the surface text level, e.g., sentence, words, word-parts, tokens, etc. Such commonly-used measures include resemblance in vocabulary distribution (Kilgarriff, 2001), likelihood of repetition (Pillutla et al., 2021), and n-gram matching (Papineni et al., 2002). However, these metrics tend to be overly sensitive or easily misled by adversarial samples or text peculiarities. In general, χ^2-based metrics calculate distance between observed and expected frequencies of categorical variables. The metric in (Kilgarriff, 2001), denoted here as **CHI**, calculates E, the average (between P and Q) frequencies of the n most common tokens in the combined vocabulary of P and Q, then sums the χ^2 statistics comparing each of P and Q to the expected E, across tokens. Here, for both CHI and ZIPF, below, we use the top $n = 5000$ tokens.

In contrast, the **ZIPF** metric (Holtzman et al., 2019) compares the use of vocabulary using Zipf's law, which suggests that the frequency of a given word in human text is inversely-proportional to its frequency rank. The Zipfian coefficient is fitted on a given corpus and the further it is from 1, the more the observed corpus differs from the 'ideal' theoretical distribution (Holtzman et al., 2019). We can thus use $|z_P - z_Q|$ as a distance metric between corpora P and Q.

Distributional Metrics These metrics are based on quantifying the distributional relationship between the reference and target corpora in the embedded vector space, thereby capturing semantics beyond superficial token-level statistics. Here P and Q denote the reference and target corpora in the embedding space. Given samples from these, we can use the sample density estimates $\hat{P}$ and $\hat{Q}$ to approximate the true unknown corpus population distributions P and Q.

[2]https://github.com/IBM/comparing-corpora

Type	Metric	Measures
Lexicographical Statistics	CHI (χ^2) (Kilgarriff, 2001)	Word/Token count comparison.
	ZIPF (Holtzman et al., 2019)	Unigram rank-frequency statistics.
Distributional	FID (Heusel et al., 2017)	Wasserstein distance between densities.
	PR (Sajjadi et al., 2018)	Assessing distributional precision & recall.
	DC (Naeem et al., 2020)	Estimating manifolds density and coverage.
	MAUVE (Pillutla et al., 2021)	Quality & diversity via divergence frontiers.
Discriminative	CLASSIFIER (2016)	Classifiability between reference and target.
	IRPR (Zhao et al., 2017)	Average distance between closest samples pairs.

Table 1: Summary of investigated text similarity metrics.

The Fréchet Inception Distance (**FID**, Heusel et al. 2017) is computed by fitting a continuous multivariate Gaussian to the P and Q, and then calculating the Wasserstein-2 distance between them. However, FID is sensitive to both the addition of spurious modes as well as to mode dropping (Lucic et al., 2018). Also, while FID is able to detect distributional distances in the high-dimensional space, it cannot shed light upon the nature of this distance.

Due to these weaknesses of FID, we additionally consider a metric denoted **PR** proposed in computer vision (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), which is inspired by the notion of precision and recall in machine Learning. Intuitively, the precision captures the average resemblance of the individual target samples to the reference set (i.e., fidelity), while the recall measures how well the target samples "cover" the full variability of the reference samples (i.e., diversity). To obtain a single distance value using the method in (Kynkäänniemi et al., 2019), we calculate the $F1$ measure based on the returned precision and recall, denoted here by PR.

Naeem et al. (2020) proposed an improved estimation of these precision and recall notions (called, density and coverage) by mitigating the overestimation of manifolds caused by outliers and underestimating the similarity when the target and reference are taken from the same distribution. Similarly to PR, we calculate the $F1$ to obtain a similarity value using this method, denoted as **DC**[3].

MAUVE (Pillutla et al., 2021) is a recently-developed metric that estimates the gap between human and generated text by computing the area under the information divergence frontiers in a quantized embedding space using the KL-divergence[4].

Discriminatability Metrics Similar to the distributional metrics, discriminative metrics calculate the distance using the embedding of the individual sentences in the two corpora. However, they do not aim to specifically capture the overlap between the distribution induced by the compared corpora. Rather, they calculate the relationship in classification terms, i.e., to what extent can sentences in one corpus be distinguished from the sentences in the other corpus, using a discriminative model.

CLASSIFIER: Following (Lopez-Paz and Oquab, 2016), we measure the similarity between corpora using a binary classifier. We used SVM (Cortes and Vapnik, 1995) trained on samples of both source corpora to predict corpus membership in a test set of unseen samples. A higher test accuracy indicates higher inter-corpora distance.

While CLASSIFIER is a model-based metric that uses the entire corpus distribution, **IRPR** (information-retrieval precision and recall) is an example of an instance- (individual sentence) based corpus distance metric. Inspired by Zhao et al. (2017), we calculate the dissimilarity between corpora as follows. For each embedded sentence in corpus A, we find its closest neighbor in B by cosine distance. The average of these distances is then computed to find the "precision" value. The same procedure in reverse, from B to A, gives the "recall" value. We calculate the $F1$ score of the recall and precision to obtain a single value. Note that the CLASSIFIER metric is used to represent model-based discriminative approaches, while IRPR is used to represent instance-based discriminative methods.

The values calculated by CHI, IRPR, PR, DC and Mauve capture the similarity rather than the distance between two corpora (for all metrics $v \in [0, 1]$). To make these metrics represent distances, we take $1 - v$.

Our model selection was based on considering the

[3]To calculate both PR and DC, we employed the implementation provided in the *prdc* Python package.

[4]We used the *mauve-text* Python package for calculating MAUVE as well as ZIPF.

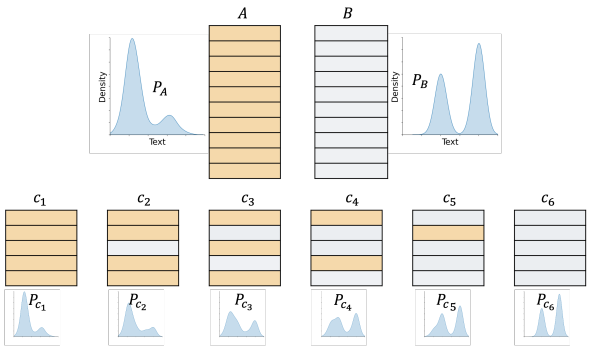

Figure 1: Construction of a $k = 6$ known similarity corpora (KSC) collection from source corpora A and B. The corpus c_i is constructed by drawing $n\left(\frac{k-i}{k-1}\right)$ and $n\left(\frac{i-1}{k-1}\right)$ samples from A and B, respectively. The adjacent densities illustrate the text distributions in the semantic space of the source and the KSC corpora.

trade-off between embedding quality and calculation time. The code as well as the scripts to reproduce the experiments are available online.[5]

3 Known Similarity Corpora

Most of the metric quality measures we propose are primarily based on the notion of *known-similarity corpora* (KSC) introduced by Kilgarriff (2001). The KSC set is created by mixing samples from two different source corpora A and B in gradually-changing proportions. The KSC set, denoted $KSC(A, B)$, consists of k corpora $\{c_1, c_2, \ldots, c_k\}$, each of size $n \geq k-1$, where corpus c_i, $i = 1, \ldots, k$ is constructed by sampling $n\left(\frac{k-i}{k-1}\right)$ observations from A, and the remaining $n\left(\frac{i-1}{k-1}\right)$ from B (see Figure 1). The sampling resolution gradation between corpora is a fixed $\frac{1}{k-1}$.

We now introduce some notation on the KSC set, which are used to define the measures in Section 4. Let $[k] = \{1, 2, \ldots, k\}$. For given source corpora A and B, for each $\ell \in [k-1]$ we define the **ℓ-distant corpora set** as follows:

$$KSC_\ell(A, B) = \{(c_i, c_j) : i, j \in [k],\ j - i = \ell\} \quad (1)$$

Let $d(a, b)$ denote the distance from corpus a to b, according to metric d. Let $D_\ell(A, B, d)$— D_ℓ for short—be the set of values of distance d for corpora pairs in $KSC_\ell(A, B)$;

$$D_\ell(A, B, d) = \{d(c_i, c_j) : (c_i, c_j) \in KSC_\ell(A, B)\} \quad (2)$$

To pool results across ℓ, we further define:

$$D(A, B, d) = \{D_\ell(A, B, d) : \ell \in [k-1]\} \quad (3)$$

[5] https://github.com/IBM/meme

Name	Size	Description
atis	4978	Utterances to a flight booking system.
yahoo	20000	Yahoo non-factoid questions in 21 categories.
clinc150	22500	Utterances in 10 domains classified into 150 classes.
banking77	10000	Online banking queries.

Table 2: Datasets used as source corpora in our benchmark. Although some of the datasets are partitioned annotated with labels, in our experiments, if not mentioned otherwise, we ignored those labels.

Some of the metrics d have a pre-defined range (e.g., CHI, MAUVE, DC, PR only return values in the range [0,1]) while others have no preset scale or operation range. Therefore, to allow sensible comparison of distance metrics with different operation ranges and across source corpora, we obtain z-scores by normalizing the metric values, pooled across all $D_\ell(A, B, d)$. In the following analysis, if not specified otherwise, D_ℓ will always be the normalized rather than raw distances.

Datasets Selection The measures described in Section 4 are applicable to any pair of textual datasets with differently-distributed textual content, allowing the corpora in the KSC set to be distinguishable. To ensure that each pair of source corpora were in fact different enough, in the following experiments we use pairs of human text corpora from different domains, rather than pairing a human corpus with a machine-generated version of itself. For our experiments we selected four public datasets (ATIS, Hemphill et al. 1990; yahoo[6]; banking77, Casanueva et al. 2020; clinc150, Larson et al. 2019) containing short user utterances from different domains summarized in Table 2.

4 Metric Robustness Measures

We now describe our measurements of desirable properties for distance metrics, given the normalized D_ℓ on the KSC sets. In the three following measures (Monotonicity, Separability, and Linearity), we aim to capture three attributes of well-behaved metrics that can be understood by considering the top line scatter plots of Figure 2; these show the relation between the D_ℓ sets and ℓ. In these scatterplots, a high angle of the regression line, low vertical variability around it, and linearity

[6] https://ciir.cs.umass.edu/downloads/nfL6/

are all desirable properties for the distance metric, and are captured in these measures.

4.1 Metric Monotonicity

A well-behaved distance metric d should have a natural monotonic relationship with the separation levels ℓ of the KSC. We use Spearman's rank correlation between ℓ and D_ℓ, which we denote $\rho(d)$, to assess the monotonicity. Spearman's correlation is defined as the Pearson correlation between the order ranks of two variables, and measures the strength of their monotonic, rather than linear, association. As can be seen in Table 3, MAUVE and CHI achieve the best monotonicity results, followed by DC and FID.

4.2 Metric Separability

It is desirable that (1) for a given ℓ, D_ℓ has low variability, and (2) for different $\ell_2 > \ell_1$, the samples D_{ℓ_1} and D_{ℓ_2} are distinguishable (e.g., by a two-sample difference test), particularly as $\ell_2 - \ell_1$ grows. Here, we measure how grouping by ℓ explains the variability in D_ℓ across ℓ. We perform a one-way fixed-effects analysis of variance (ANOVA) with ℓ as the unordered categorical treatment and D_ℓ as the numeric response. Often, an F-test is performed; if its p-value is low, it means a significant amount of the variance in the response (D_ℓ) can be explained by the treatment (ℓ). Since the F-test for any reasonable d metric should be significant, we instead use the similar ω^2 effect-size metric (Hays, 1963), which is bounded by ± 1, to better assess them. It is defined as

$$\omega^2 = \frac{SS_{\text{treatment}} - df_{\text{treatment}} \times MS_{\text{residual}}}{SS_{\text{total}} + MS_{\text{residual}}} \quad (4)$$

where SS and MS are the sum and mean sums of squares, and df is the degrees of freedom, on a dataset of size n (here, $n = |D(A, B, d)|$). In the following we denote this measure as $\mathcal{W}(d)$.

4.3 Metric Linearity

Here we examine to what extent linear changes in the corpus content (ℓ) are manifested in linear changes in the distance function. To do so, we calculate the coefficient of determination (R^2), where higher values indicate stronger linearity. This measure is denoted by $\mathcal{L}(d)$. Looking at the results in Table 3, we see that MAUVE achieves the best results followed by DC and FID. It appears that this measure is more affected by the source corpora and by the resolution than other metrics.

4.4 Metric Time Efficiency

The time complexity of the metric is commonly perceived as less important, thus seldom reported (Sai et al., 2022). This aspect is becoming ever more important, especially due to the growing interest in time-consuming divergence frontier methods (Djolonga et al., 2020). Such metrics perform multiple measurements to estimate the area under the curve (similar to precision-recall curves for binary classification), with tune-able but increasing resolution. We measure the time performance of the metric $\mathcal{T}(d)$ in terms of 100 similarity measurements operations per second ($[100op/sec]$) on a standard CPU machine[7]. Note that the measurements reported in Table 3 do not include the sentences' embedding time. Predictably, methods that operate on the token level and avoid complex density estimation tend to achieve the best time performance. Among the distributional metrics, MAUVE achieves the best results, followed by FID. PR and DC produce similar results since both are based on similar manifold calculations.

4.5 Metric Accuracy

The assessment measures described earlier in Section 4 use the observed values of the metric distances (or similarities) between the KSC corpora; however, the actual values of the distance may not be known. Nevertheless, we still have some partial information about the ordering of these values, which we will use to define an accuracy measure.

4.5.1 Comparing paired corpora distances

Even though we do not have the true distance between any two corpora in the KSC set, we can still assume that certain pairwise distances are larger than others. For instance, it should be true that, say, $d(c_2, c_3) \leq (c_1, c_4)$ in expectation (across repeated random sampling). This is because the proportions of observations from A in c_2 and c_3 are more similar than the respective proportions between c_1 and c_4. Moreover, the interval of the first pair is 'contained'[8] in the second, and thus the first pair should have smaller distance. Thus, In general, whenever the interval of one corpus pair contains ($\subset$) the interval of another, we expect the contained pair to have a smaller distance.

[7] CPU: 2.3 GHz 8-Core Intel Core i9. Memory: 64 GB DDR4 (2667 MHz)

[8] (c_i, c_j) contains (c_q, c_r), i.e., $(q, r) \subset (i, j)$, if $i \leq q$ and $r \leq j$ and $i < r$.

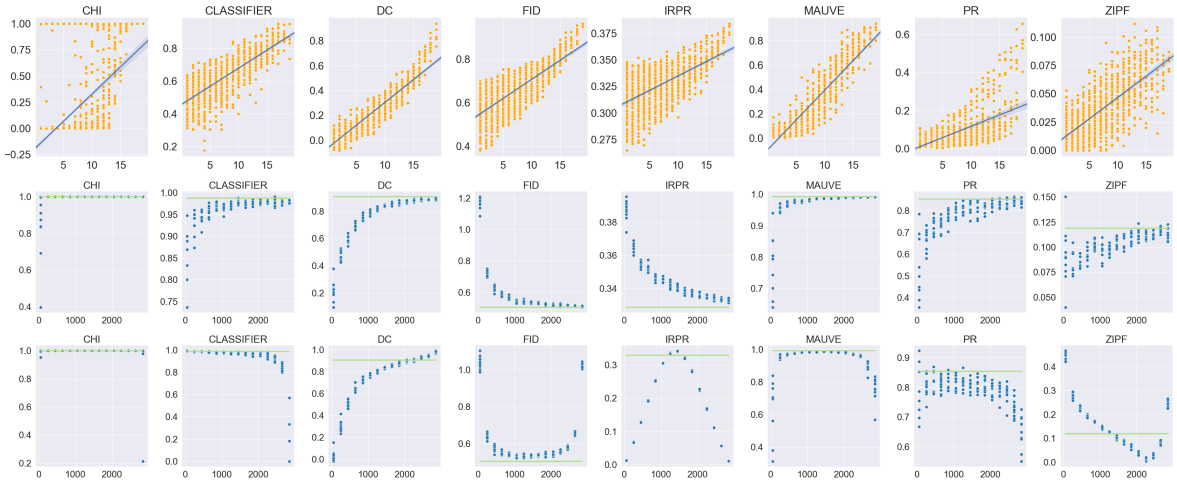

Figure 2: Top: Distance values (non-normalized) of corpora pairs in D_ℓ versus ℓ. ($n = 100, k = 12, |J| = 6053$), pooled across 5 repetitions of KSC samples. Blue line indicates regression and confidence interval at 95%. Middle: Distance values calculated on increasing s size corpora a_s and b_s sampled from sources A and B, correspondingly. Bottom: Distance between imbalanced corpora a_s and $b_{\bar{s}}$ where $|b_{\bar{s}}| = N - s$ and $N = 2900$. The x-axis represents $s \in \{50, 250, 450, \ldots, 2850\}$ ($repetitions = 10$). In middle and bottom figures, green horizontal line indicates the asymptotic distance $d(A, B)$. In all figures A=clinc150 and B=banking77.

	$\mathcal{A}(d)$		$\mathcal{A}^w(d)$		$\mathcal{T}(d)$		$\rho(d)$		$\mathcal{W}(d)$		$\mathcal{L}(d)$		$\mathcal{S}(d)$	$\mathcal{I}(d)$
k	7	12	7	12	7	12	7	12	7	12	7	12		
CHI	.945	.852	.913	.774	**4.68**	**3.29**	.875	.866	.684	.702	.810	.767	**.989**	**.994**
CLS.	.789	.701	.731	.618	1.26	1.10	.704	.735	.544	.562	.767	.767	.972	.918
DC	.958	.863	.936	.805	.031	.031	.913	**.892**	**.908**	**.879**	**.946**	**.919**	.832	.808
FID	.949	.810	.923	.753	.067	.066	.764	.695	.563	.537	.81	.759	.821	.877
IRPR	.832	.710	.784	.638	4.39	2.35	.571	.543	.258	.275	.645	.598	.949	.856
MUV.	**.976**	**.888**	**.963**	**.828**	.079	.071	**.938**	**.906**	.883	**.885**	**.947**	**.926**	**.977**	.943
PR	.820	.688	.767	.608	.031	.031	.649	.592	.577	.566	.716	.667	.909	.934
ZIPF	.886	.726	.851	.657	4.65	2.668	.751	.633	.514	.413	.785	.667	.852	.913
CHI	.953	.935	.936	.891	**5.58**	**3.33**	**.960**	**.962**	.835	**.900**	.83	.858	**1.00**	**1.00**
CLS.	.931	.827	.902	.751	1.29	1.21	.843	.836	.671	.702	.847	.847	**.993**	**.989**
DC	.773	.601	.702	.552	.031	.031	.707	.615	.763	.717	.825	.759	**.988**	**.986**
FID	.967	.904	.947	.848	.071	.067	.793	.754	.634	.636	.845	.816	.922	.898
IRPR	.697	.587	.642	.570	3.91	2.69	.382	.264	-.02	-.001	.433	.341	.951	.834
MUV.	**.999**	**.977**	**.998**	**.961**	.084	.067	.936	.943	**.856**	**.904**	**.932**	**.950**	**.999**	**.994**
PR	.722	.467	.666	.446	.031	.030	.459	.240	.488	.394	.658	.523	.890	.899
ZIPF	.854	.783	.817	.736	4.77	3.02	.660	.635	.309	.352	.687	.661	.735	.904

Table 3: Summary of metrics evaluation scoring on two pairs of source datasets in low ($k = 7$) and high ($k = 12$) resolution KSC ($n = 100$). Best results with differences below .015 are marked in bold. $\mathcal{T}(d)$ units are $[100op/sec]$. MUV. stands for MAUVE and CLS. for CLASSIFIER. In the top table, A=clinc150 and B=banking77. In the bottom table A=atis and B=yahoo. The average results of 5 repetitions are reported for all measures except size and imbalance robustness, in which the number of repetitions is 10. More statistical details are provided in Figure 6 in the Appendix.

Given two pairs, (c_i, c_j) and (c_q, c_r), of paired corpora, we can only reliably predict which of $d(c_i, c_j)$ or $d(c_q, c_r)$ is larger in expectation (a decision we call a 'judgement') if the interval of one pair contains the other's. The set J contains all and only such judgements:

$$J = \{((c_q, c_r), (c_i, c_j)) : (q, r) \subset (i, j)\} \quad (5)$$

The judgement $d(c_q, c_r) \leq d(c_i, c_j)$ is correct when the second interval contains the first. This gives the most probabilistically-logical partial order on the similarities between corpora in a KSC collection, that can be obtained without knowledge of the true pairwise d-distances between corpora[9]. Figure 3 shows a tree representation of KSC-set pair containment relations, from which the set of judgements J can be extracted.

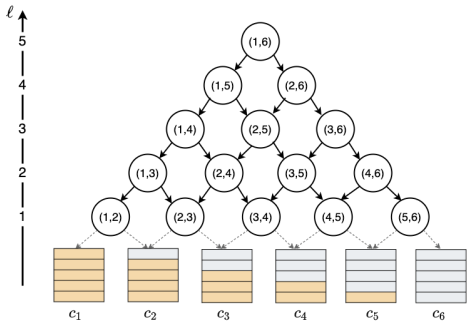

Figure 3: A tree representation of the judgements performed on the KSC collection given a metric $d(\cdot, \cdot)$, for calculating the accuracy ($\mathcal{A}$, Section 4.5) measures. The leafs are the KSC collection and the inner nodes (circles) represent the corpora tuples (c_i, c_j). The set J contains all judgements such that each node (i, j) is judged against all descended nodes. Namely, if there is a path from node a to node b, there is a judgement between the two nodes, and the judgement is correct if $d(b) \leq d(a)$. The size of the judgements set can be expressed as: $|J| = \sum_{i=1}^{k}(k-i)\left(\frac{i(i-1)}{2} - 1\right)$. For instance, $|J| = 339$ if $k = 7$, and 6053 if $k = 12$.

4.5.2 Accuracy

The metric accuracy is defined as the rate of correct judgements, formally:

$$\mathcal{A}(d) = \frac{1}{|J|} \sum_{\jmath \in J} \mathbb{1}(d(c_q, c_r) \leq d(c_i, c_j)) \quad (6)$$

[9]For instance, say we compare pairwise distances between (c_1, c_6) and (c_5, c_7). Even though the second interval length $(7 - 5 = 2)$ is smaller than the first $(6 - 1 = 5)$, because it is not contained in the first, we cannot necessarily say that $d(c_5, c_7) \leq d(c_1, c_6)$ since inter-corpus distance may not be proportional to the interval length.

where $\jmath = ((c_q, c_r), (c_i, c_j))$ is a judgement in J and $\mathbb{1}(\cdot)$ is the indicator function. Further, we propose a weighted version of the accuracy metric that assigns more weight to harder judgements. We define the hardness of judgement $\jmath$ as $w(\jmath) = \frac{1}{\ell_2 - \ell_1}$ where $\ell_2 = j - i$ and $\ell_1 = r - q$, and $\ell_2 > \ell_1$ by definition of J. Formally,

$$\mathcal{A}^w(d) = C \sum_{\jmath \in J} w(\jmath) \cdot \mathbb{1}(d(c_q, c_r) \leq d(c_i, c_j))) \quad (7)$$

where $C = \left(|J| \cdot \sum_{\jmath \in J} w(\jmath)\right)^{-1}$. While $\mathcal{A}$ and $\mathcal{A}^w$ are correlated, as one may expect, $\mathcal{A}^w$ typically returns lower values (see Table 3).

In our implementation, the set of samples in each c_i is disjoint, namely, $c_i \cap c_j = \varnothing, \forall c_i, c_j \in KSC(A, B)$. This was done to prevent perfect judgements by naively counting the number of common instances (e.g., by defining $d(c_i, c_j) = |c_i \Delta c_j|$ where Δ denotes the symmetric difference). MAUVE, followed closely by FID, CHI and DC, achieves the highest accuracy results across resolutions and source corpora.

4.6 Size Robustness

We are also interested in capturing the sensitivity of a metric to sample sizes. To accomplish this, we need to quantify the convergence pace of $d(a_s, b_s)$ to the asymptotic distance $d(A, B)$, where a_s, b_s are samples from corpora A, B of increasing size s. Specifically, in our experiments $s \in S = \{50, 250, 450, \ldots, 2850\}$. The middle plot in Figure 2 shows convergence patterns of the different metrics to the asymptotic distance. The asymptotic distance is estimated by the mean of repeated (10) calculations of the distance on samples of size 3000 each from A, B, rather than on the full corpora. To quantify the metric size robustness, (S), we calculate the mean absolute error, $|d(a_s, b_s) - d(A, B)|$, for all $s \in S$, normalized by the asymptotic distance:

$$\mathcal{S}(d) = 1 - \sum_{s \in S} \frac{|d(a_s, b_s) - d(A, B)|}{d(A, B)} \quad (8)$$

Similar to previous measures, the normalization is performed to omit the influence of metric scale and operation ranges.

While our results demonstrate (Figure 2) that most of the metrics examined require around 1000 samples to closely estimate the asymptotic distance between the source corpora, their measured accuracy ($\mathcal{A}$(d) and $\mathcal{A}^w(d)$) is still fairly high even on

small corpora within the KSC, and can capture relative differences in corpus content.

4.7 Imbalance Robustness

Similarity metrics are frequently used to compare datasets with unequal sample size. Especially when comparing real and generated corpora, the size of a generated corpus is usually much larger than the real corpus. The imbalance robustness measure quantifies the effect of corpora size imbalance on the metric's performance (see Figure 2, bottom).

Unsurprisingly, asymmetric metrics such as PR and DC are most affected by size imbalance. While PR, DC, and MAUVE were all originally designed to measure the disparity between human and generated data (and thus asymmetric in the reference$P$ and target Q), it seems that MAUVE overcomes the sensitivity to datasets of very unequal sizes. Interestingly, imbalance causs some metrics (CLASSIFIER and MAUVE) to underestimate the distance, while others (FID) overestimate it. When we compare the convergence patterns of PR and DC, both are similarly asymmetric, maintaining $d(P, Q)$. When we increase the reference size, PR diverges from the true asymptotic distance, while DC converges to it. The Imbalance Robustness score $\mathcal{I}(d)$ is calculated similarly to the size robustness score, only that $|b_s| = N - |a_s|$.

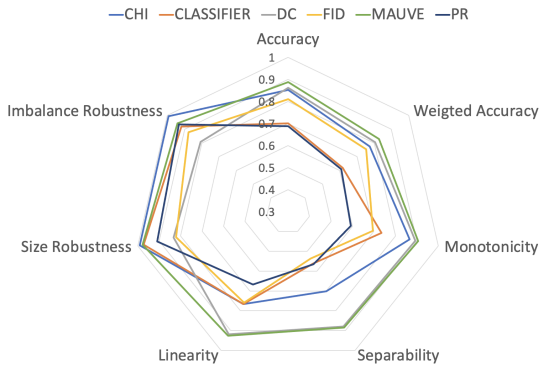

Figure 4: Leading metrics characterization radar chart. Mean results from Table 3 for A=clinc150, B=booking77 and for $k = 12$, excluding time efficiency to maintain scale.

KSC Parameters As shown in Section 4.6, most metrics require at least $n = 1000$ samples to capture the true distance between two source domain corpora; however, our experiments use $n = 100$. This is because our measures are relative, i.e., we do not aim to calculate the true asymptotic distance between two domains, but to measure the metrics' robustness in detecting small changes in the compared corpora. Furthermore, if n is large, k must also be large to ensure the k corpora in the KSC set have small enough absolute consecutive differences.

Note that small consecutive differences in KSC corpora are needed so that the measures in Section 4 will have a high enough resolution and large enough sample size of D_ℓ to properly differentiate the metric properties. In particular, this ensures the judgements (Section 4.5.1) used in the accuracy measures ($\mathcal{A}$ and $\mathcal{A}^w$) are not too 'easy' to make correctly, in which case they would be less useful as a tool. For instance, a metric with 100% accuracy makes all correct judgements, e.g., that $d(c_2, c_3) \leq d(c_1, c_4)$. If $k = 5$, the gap (in expectation) between the pair distances compared is large, so the judgement is easy, and thus all metrics may have full accuracy. When k increases, the absolute consecutive differences in corpora fall, and thus the difficulty of the judgement increases. Some metrics will fail to make the judgement correctly (in a given random KSC), decreasing their accuracy; this allows us to better differentiate between the more and less accurate metrics. However, setting k too high results in a computationally prohibitive number $|J|$ of judgements. Therefore, we opted to use the smaller n that are still sufficient to capture the quality and robustness of the investigated metrics.

5 Increasingly Fine-tuned Corpora

Here, we qualitatively investigate the metrics' ability to discriminate between generated and human text using the following procedure: We generated a sequence of equal-size synthetic corpora $IFC = (g_1, g_2, \ldots, g_n)$ by sampling from a gradually fine-tuned language model on a specific source corpus A. Namely, in each iteration, a fine-tuning step is performed by training the language model on a single epoch containing 1000 sentences randomly drawn from A, followed by a generation process to synthesize a corpus g_i containing around 1000 sentences. The name IFC, or "Increasingly Fine-tuned Corpora", was chosen to parallel the name KSC ("Known Similarity Corpora").

For each generated corpus g_i, we estimated the distance from A, i..e, $d_i = d(A, g_i)$, $\forall i \in [n]$. While the true distance between those synthetic corpora and A are unknown, an effective metric should

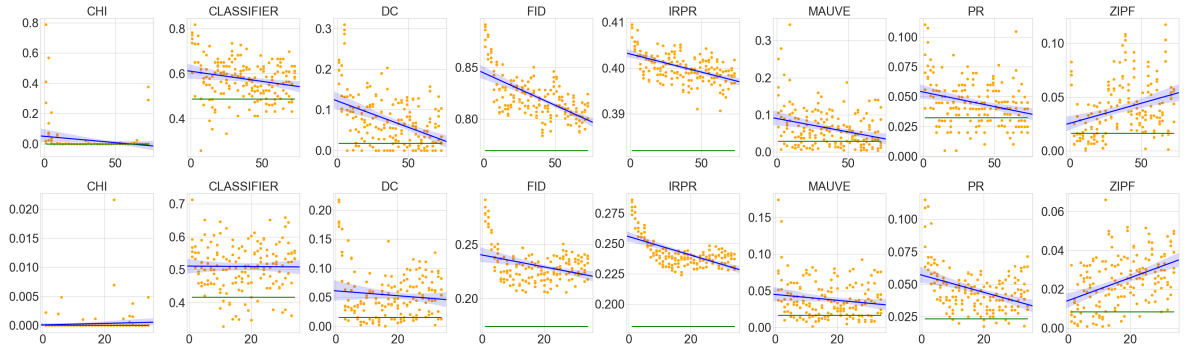

Figure 5: Similarity between reference corpus and iteratively fine-tuned corpora g_i samples. Orange dots show the similarity between samples of generated text in iteration i and the source dataset. The blue line indicates regression and confidence interval at 95%. The green horizontal line specifies the mean estimation of the distance between two random samples of the original corpus. The top figure shows iterative generation on unlabeled news headlines dataset. The bottom shows the iterative conditional generation using LAMBADA (Anaby-Tavor et al.) trained on banking77 dataset.

capture the decreasing distance between A and g_i with increasing i, namely $d_1 \prec d_2 \prec \cdots \prec d_n$. Due to our results, which show low imbalance robustness of some metrics, we maintained the same-size corpora when calculating corpora distance.

The results presented in Figure 5 show the gap between human and generated text captured by each metric in each iteration. To calculate the average self-distance of the reference corpus (A), we take the mean distance between two randomly sampled sub-corpora r_1 and r_2 from A, i.e. $d(A, A) = \mathbb{E}_{r_1, r_2 \sim A}[d(r_1, r_2)]$).

In our experiments we used two datasets, the banking77 dataset, mentioned above and the news dataset[10], representing different domains of text corpora. The IFC set for the banking77 dataset was generated in an iterative two-step procedure similar to the one described in LAMBADA (Anaby-Tavor et al.). This procedure first generates sentences conditioned on the label, then filters out sentences that are out-of-domain or incorrectly labeled. However, the IFC set for the news dataset was generated by finetuning the pre-trained GPT-2 medium model (Radford et al., 2019).

The results in Figure 5 show that CHI is less effective than the other metrics in capturing the gradual nature of the IFC. Also, they show that FID and IRPR are sensitive in discriminating between the original and generated corpora, even after many fine-tuning iterations. Interestingly, the ZIPF distance increases with the iteration. This indicates that the generated text, despite becoming semantically closer to the original with the increasing iterations, becomes less 'natural' in that the token frequencies deviate from that of human text and the reference corpus. This can be explained, at least in part, by the TTR measure. TTR is a standard word diversity measure, calculated by dividing the number of unique words in a text by the total word count. A high TTR indicates significant lexical variation. Indeed, in the IFC of banking77, g_1's TTR is 0.295 which is closer to the original dataset's TTR of 0.299 than g_{40}'s TTR of 0.322.

6 Conclusions

In this work, we propose a principled set of automatic measures for evaluating the quality of text dissimilarity metrics. By testing various metrics using our measures, we show that they do a good job of capturing their known characteristics, hence increasing our confidence in these measures; also, overall, recent metrics exhibit more favorable traits than their predecessors. The radar chart in Figure 4 shows that our measure scores correlate well with the compared distributional metrics recency $MAUVE \succ DC \succ PR \sim FID$ as well as their known relative strengths.

7 Limitations

Although one of the main motivations for comparing corpora is to measure the semantic gap between human and generated short text, we used pairs of human text corpora from different domains to maintain controllably-distinct corpora in the KSC set. Despite this, future efforts to develop human and machine-generated benchmark pairs (Mille et al., 2021) will allow for future work to quantitatively

[10] HuffPost (www.huffpost.com) news headlines collected from 2012 to 2018 containing around 200k headlines.(www.kaggle.com/rmisra/news-category-dataset (https://www.huffpost.com)

measure the characteristics of semantic metrics on pairs of human and generated corpora using the approach devised in this paper.

Also, for more straightforward comparisons, we used only a single sentence-embedding model. However, as other studies (e.g., GPT-2 (Radford et al., 2019) in Pillutla et al. (2021) and Bert (Devlin et al., 2018) in Lo (2019)) have shown, the quality of a corpus distance metric can be affected by the embedding choice. In future extensions of our work, we plan to allow for multiple embeddings to obtain a more refined evaluation of the metrics.

An important limitation of this work is that it considers only English corpora of short text samples. We examined only a limited set of metrics and datasets, both of which we intend to extend.

In addition, we note that while our experiments calculate all KSC-based measures using a single KSC collection (same n and k values), it could be favourable to use different n and k for different measures. For instance, the time performance is calculated using a single size small dataset $n = 100$. In future work, the time scalability of metrics can be more closely investigated by comparing their time performance on increasing corpora sizes.

As indicated in Section 4, creating KSC collections with large k creates an excessive number of judgements (e.g., for $k > 15$, $|J| > 50000$), thus limiting the scalability of our method to smaller k and thus smaller n, if high resolution is required. This would preclude comparing the robustness of metrics that require large samples. We intend to rectify this in future work by creating representative smaller judgement sets by carefully sampling from the complete set.

As mentioned in Section 2, some of the investigated metrics were adapted to return a single value summarizing the distance between two corpora (e.g., averaging the precision and recall by the $F1$ score). Further work is required to build measures that can compare metrics returning multiple values.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. *CoRR*, abs/2003.04807. https://huggingface.co/datasets/banking77.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

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*.

Josip Djolonga, Mario Lucic, Marco Cuturi, Olivier Bachem, Olivier Bousquet, and Sylvain Gelly. 2020. Precision-recall curves using information divergence frontiers. In *International Conference on Artificial Intelligence and Statistics*, pages 2550–2559. PMLR.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina McMillan-Major, Anna Shvets, Ashish Upadhyay, Bingsheng Yao, et al. 2022. Gemv2: Multilingual nlg benchmarking in a single line of code. *arXiv preprint arXiv:2206.11249*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

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

William L. Hays. 1963. *Statistics for psychologists*. Holt, Rinehart, & Winston.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*. https://www.kaggle.com/datasets/hassanamin/atis-airlinetravelinformationsystem.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Adam Kilgarriff. 2001. Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved precision and recall metric for assessing generative

models. *Advances in Neural Information Processing Systems*, 32.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507.

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.

David Lopez-Paz and Maxime Oquab. 2016. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. 2018. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31.

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.

Simon Mille, Kaustubh D Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic construction of evaluation suites for natural language generation datasets. *arXiv preprint arXiv:2106.09069*.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. 2020. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

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.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34.

Andrei Popescu-Belis. 2003. An experiment in comparative evaluation: humans vs. computers. In *Proceedings of Machine Translation Summit IX: Papers*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *The workshop on comparing corpora*, pages 1–6.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. 2018. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31.

Jiapeng Wang and Yihong Dong. 2020. Measurement of text similarity: a survey. *Information*, 11(9):421.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.

A Appendix

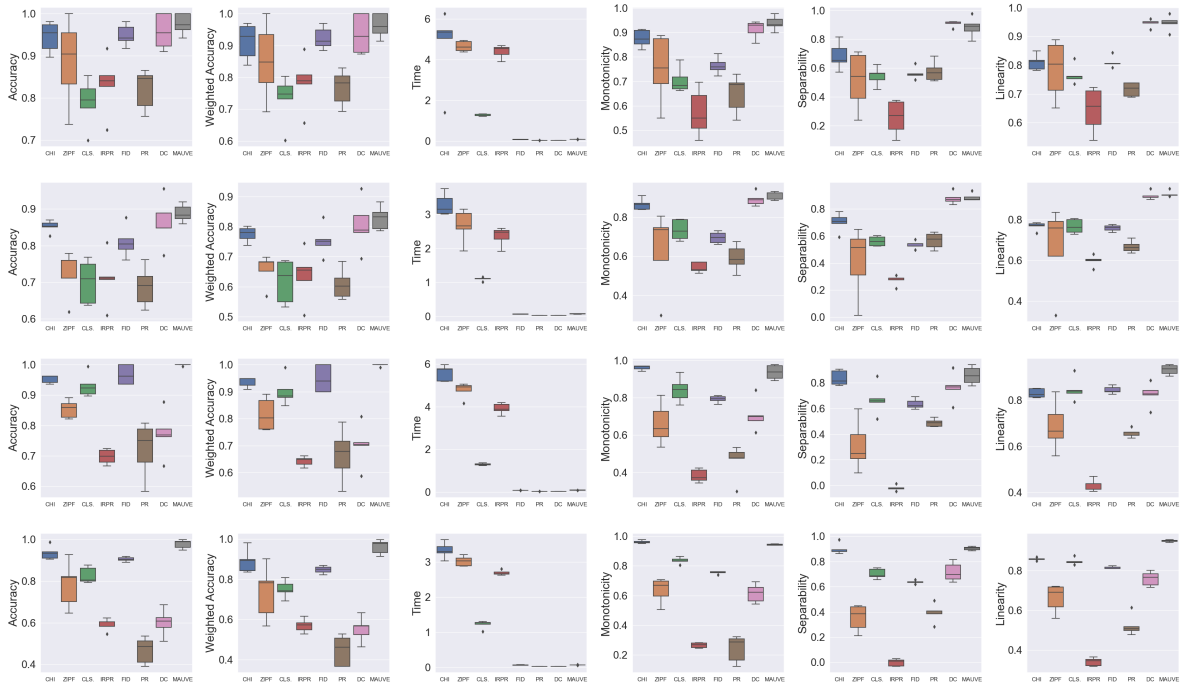

Figure 6: Distributional information of the results shown in Table 3. The top two figures showing the results for (A=clinc150 B=banking77), for k = 7 and k = 12, respectively. The bottom two figures are for (A=yahoo B=atis), for k = 7 and k = 12, respectively. Colored boxes depict the interquartile (25^{th} to 75^{th}) range. The mean is indicated by a horizontal line. All data points within 1.5 of the corresponding limits of the interquartile range are depicted by whiskers. Data points outside this range are plotted individually. CLS. indicates the CLASSIFIER metric.