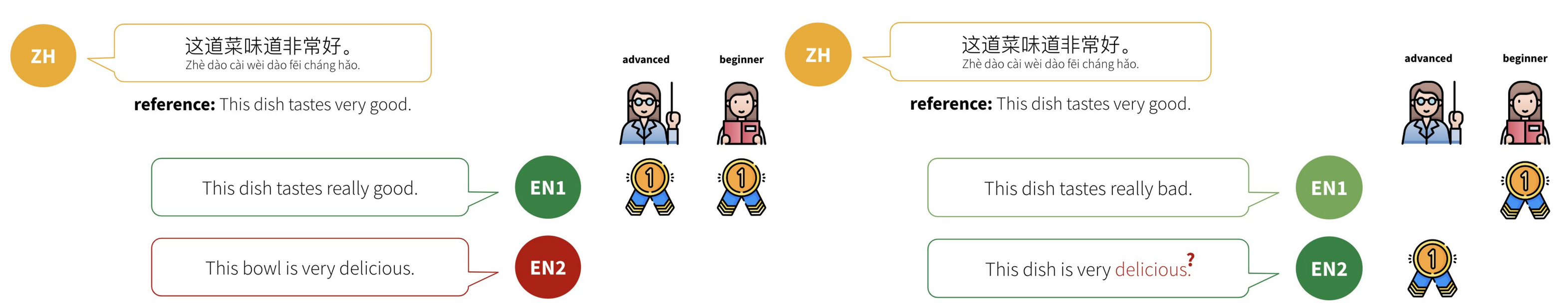
Measuring Local Accuracies to Assess Evaluation Metrics

Athiya Deviyani, Fernando Diaz
Carnegie Mellon University

Motivation

The output distribution measured by automated metrics may *shift*

- Previous research on distribution shift observe how outputs from a model change as the input distribution changes; in our case, the inputs are system outputs, the model is the metric, and the outputs are the metric scores [1]



Existing work in automated metric evaluation looks at the performance of a metric in aggregate [2][3], i.e. do not consider the fact that the performance depends on the output distribution.

Idea: a metric's ability to perform preference-based evaluation on two system outputs change as the distribution of the outputs change

- How do we measure this?

Problem Definition

Decision-level metric accuracy: for each pair of system outputs, calculate the binary difference of metric scores and the binary difference in average human judgements

- In other words, given two outputs A and B, where we know that A is objectively better than B, how often does a metric correctly assigns output A a higher score than output B?

Let $\mathscr{X}$: set of all possible system contexts
$\mathscr{Y}$: set of all possible system decisions

We define $X \subset \mathscr{X}$ to be the set of evaluation contexts
$Y_x \subset \mathscr{Y}$ as the subset of evaluation decisions $x \in X$

Assuming we have access to a perturbation function that, with high probability, degrades the utility of a decision y. Let Q_x be the set of pairs of decisions y and their corresponding degraded version y': $Q_x = \{\langle y, y' \rangle\}_{y \in Y_x}$

Let $\mu : \mathscr{X} \times \mathscr{Y} \rightarrow \Re$ be an evaluation metric that generates a scalar number reflecting the performance according to some system property that we want to measure.

Let μ^* be the ideal evaluation metric: in cases where we know that $\mu^*(x,y) > \mu^*(x,y')$, we want to observe how often $\mu(x,y) > \mu(x,y')$. This is under the assumption that μ was designed to approximate μ^*.

From the above, we formally define **local metric accuracy:**

$$\mathrm{Acc}_\mu(Q) = 1/|X| \sum_{x \in X} 1/|Q_x| \sum_{\langle y,y' \rangle \in Q_x} \mathbb{1}[\mu(x,y) > \mu(x,y')]$$

Where $Q = \bigcup_{x \in X} Q_x$

Hypothesis

	X	Y	Z
μ_A	0.9	0.8	0.7
μ_B	0.7	0.9	0.8
μ_C	0.8	0.7	0.9

Hypothesis A: the absolute local accuracy $\mathrm{Acc}_\mu(Q)$ of a metric μ changes as the subset of outputs Q changes (row-wise change)

Hypothesis B: the relative local accuracy of a metric, i.e. the total ordering of the local accuracies $\{\mathrm{Acc}_\mu(Q)\}$ of all metrics within a subset changes as the subset of outputs Q changes (cross-column change)

Methodology

Task	Dataset	Metrics
Machine Translation	System outputs and reference translations submitted to the WMT metrics task from year 2023 [4] for en-ru, en-de, and zh-en	BERT, ROUGE-1, ROUGE-2, ROUGE-L, METEOR, BertScoreP, BertScoreR, BertScoreF1, COMET, BLEURT, CHRF, UniteSRC, UniteREF, UniteUNIFIED
Automated Speech Recognition	System outputs from ESPnet models [5] on the LibriSpeech 100 dataset [6]	Word Error Rate (WER), Match Error Rate (MER), Word Information Lost (WIL), Word Information Preserved (WIP), Character Error Rate (CER)
Ranking	Ranked lists top-100 items retrieved by recommender algorithms [7] on the MovieLens1M dataset [8] submitted to TREC	Mean Average Precision (MAP), Binary Preference Score (BPREF), Precision@Relevance (RPREC), Reciprocal Rank, Interpolated Precision at Standard Recall Level@K, Precision@K

Perturbation functions to obtain y and y'

Machine Translation and Automated Speech Recognition: Remove 20% of the words in the outputs, rounded to the nearest integer [9][10][11]

Ranking: Swap the retrieval score of the items (hence swapping their corresponding rankings) within the top-100 items

- To ensure that the result of the swapping generates a random permutation, we use the following formula [12] to determine the number of transpositions k:

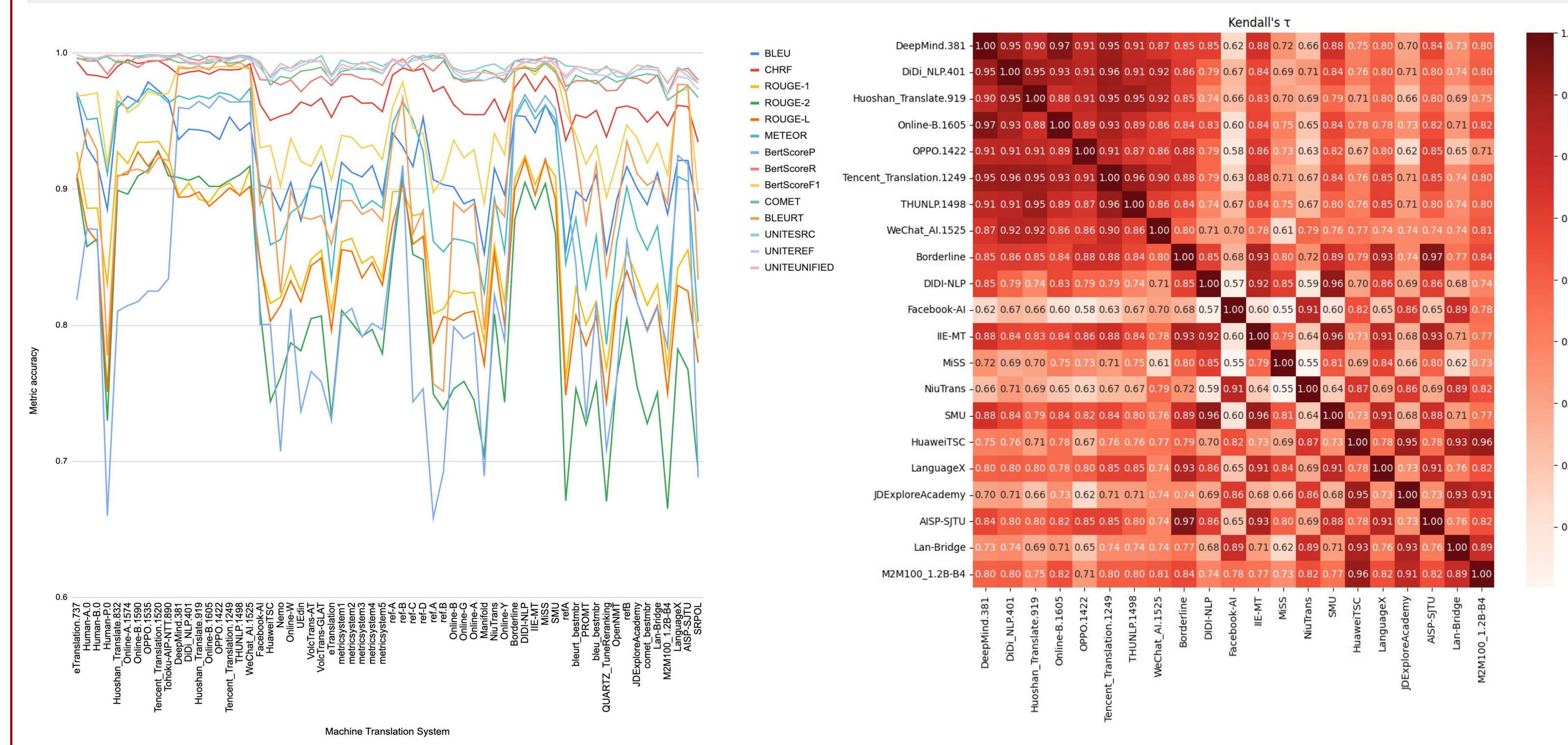
$$k = ½ * n \log(n)$$

where n is the number of items per user (our case $n = 100$); thus $k = 100$.

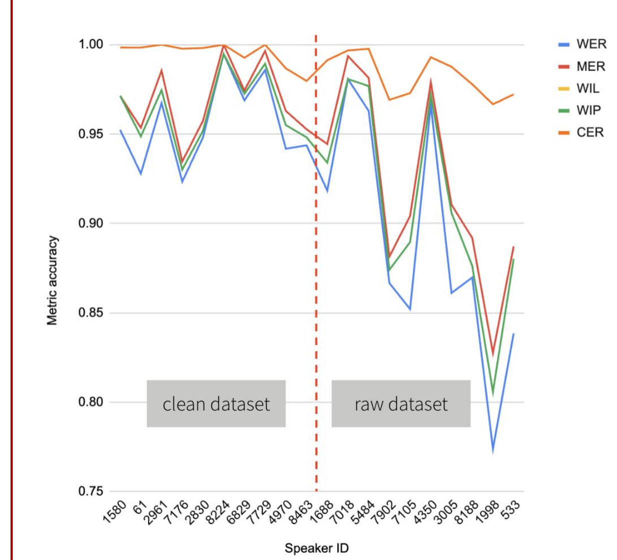
For each system output y in a dataset, we perturb them to obtain y'. Then, for each metric associated with a task, we compute how often does it correctly assigns a higher score for y than y'.

Results

Machine Translation



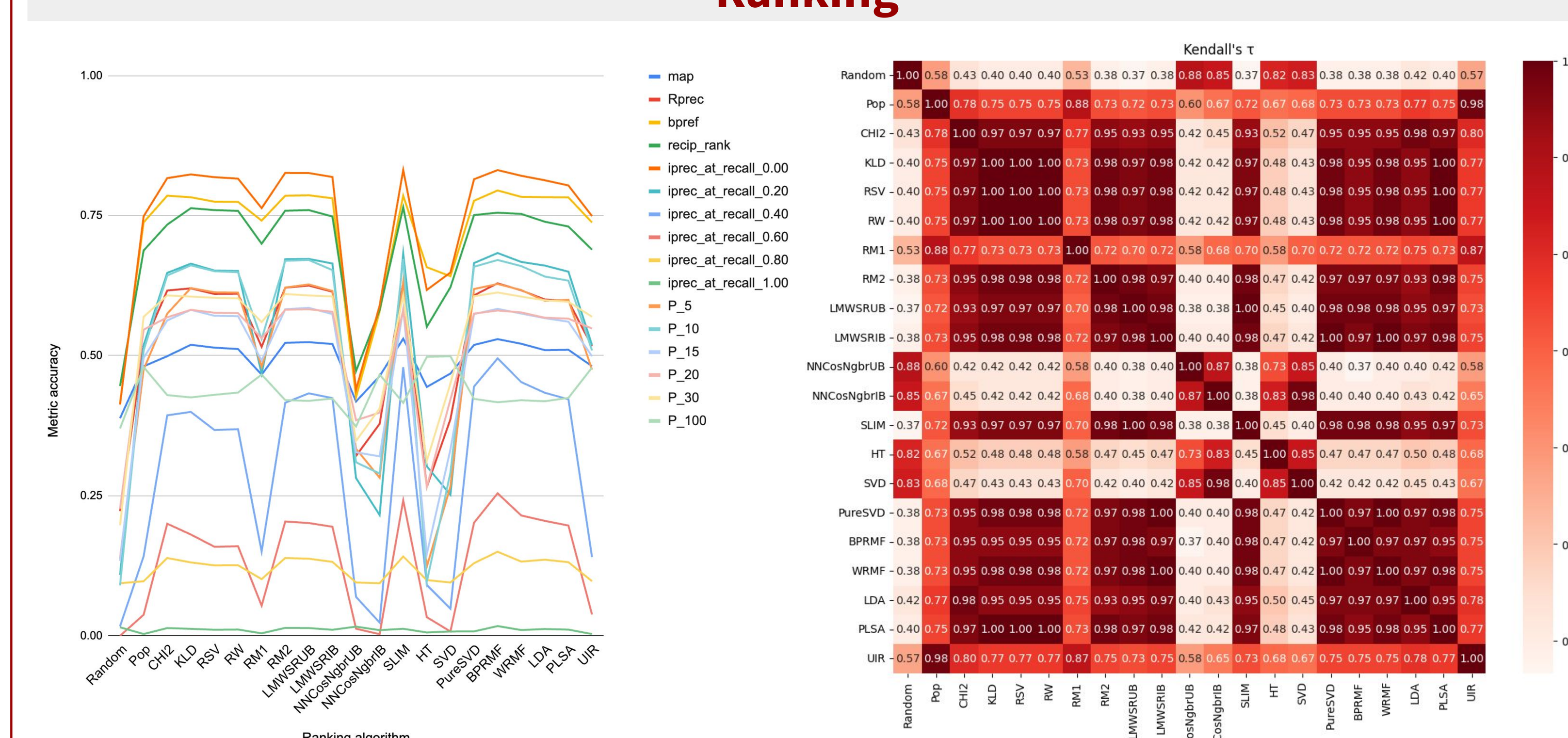
Automated Speech Recognition



We see similar supporting evidence for Hypothesis A, however, we do not see sufficient evidence that supports Hypothesis B.

Why? Metrics used in the Automated Speech Recognition task do not vary in the construct they are trying to measure and the way they are operationalized (statistical-based). Additionally, ASR is a very objective task, there are rarely multiple correct answers

Ranking



Measuring local accuracies provides a different perspective to evaluate existing evaluation metrics (it is an additional tool!)

- It is important to look at all areas in the graph, not only the metrics that has the highest accuracy at a particular subset

The value of measuring local accuracies largely depends on the nature of the task and available metrics. Based on our observation, it appears that Hypothesis A is always true, Hypothesis B is sometimes true.

[1] Yuan, Lifan, et al. "Revisiting Out-of-distribution Robustness in NLP: Benchmarks, Analysis, and LLMs Evaluations." Advances in Neural Information Processing Systems 36, 2024.
[2] Callison-Burch, Chris, et al. "Proceedings of the Third Workshop on Statistical Machine Translation." Proceedings of the Third Workshop on Statistical Machine Translation, 2008.
[3] Xiao, Ziang, et al. "Evaluating Evaluation Metrics: A Framework for Analyzing NLG Evaluation Metrics using Measurement Theory." The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.
[4] Freitag, Markus, et al. "Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent." Proceedings of the Eighth Conference on Machine Translation, 2023.
[5] Watanabe, Shinji, et al. "ESPnet: End-to-End Speech Processing Toolkit." Proceedings of Interspeech, 2018.
[6] Panayotov, Vassil, et al. "Librispeech: an asr corpus based on public domain audio books." 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015.
[7] Valcarce, Daniel, et al. "On the robustness and discriminative power of information retrieval metrics for top-N recommendation." Proceedings of the 12th ACM conference on recommender system, 2018.
[8] Harper, F. Maxwell, and Joseph A. Konstan. "The movielens datasets: History and context." ACM transactions on interactive intelligent systems (tiis) 5.4, 2015.
[9] Chen, Yanran, and Steffen Eger. "Menli: Robust evaluation metrics from natural language inference." Transactions of the Association for Computational Linguistics 11, 2023.
[10] Sai, Ananya B., et al. "Perturbation CheckLists for Evaluating NLG Evaluation Metrics." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.
[11] Chen, Anthony, et al. "Evaluating question answering evaluation." Proceedings of the 2nd workshop on machine reading for question answering, 2019.
[12] Diaconis, Persi, and Mehrdad Shahshahani. "Generating a random permutation with random transpositions." Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57.2, 1981.