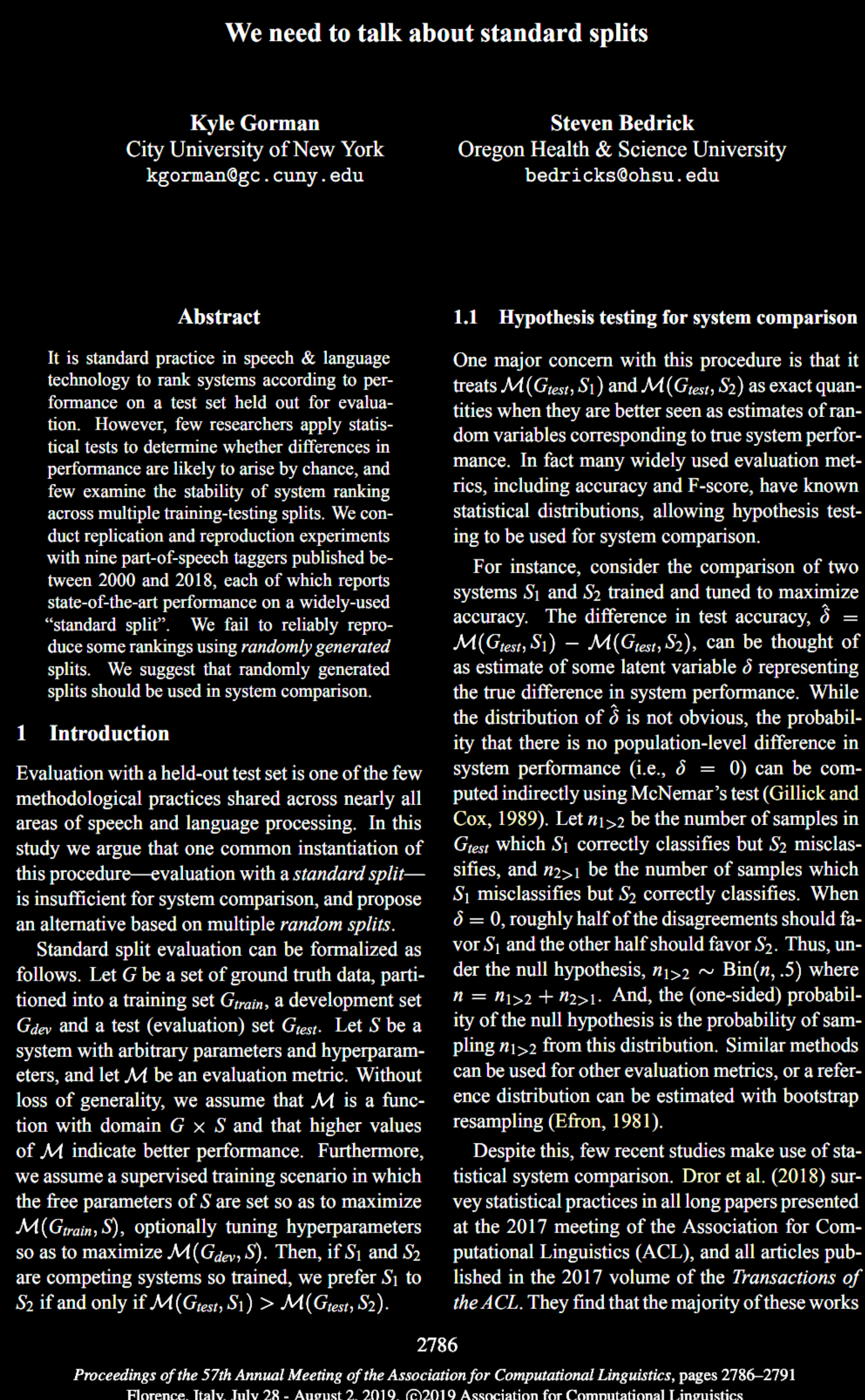
We Need to Talk About Random Splits

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, Katja Fillipova
http://anderssoegaard.github.io

coAStaL

UNIVERSITY OF COPENHAGEN

Motivation



We need to talk about standard splits

Kyle Gorman
City University of New York
kgorman@gc.cuny.edu

Steven Bedrick
Oregon Health & Science University
bedricks@ohsu.edu

Abstract

It is standard practice in speech & language technology to rank systems according to performance on a test set held out for evaluation. However, few researchers apply statistical tests to determine whether differences in performance are likely to arise by chance, and few examine the stability of system ranking across multiple training-testing splits. We conduct replication and reproduction experiments with nine part-of-speech taggers published between 2000 and 2018, each of which reports state-of-the-art performance on a widely-used "standard split". We fail to reliably reproduce some rankings using *randomly generated* splits. We suggest that randomly generated splits should be used in system comparison.

1 Introduction

Evaluation with a held-out test set is one of the few methodological practices shared across nearly all areas of speech and language processing. In this study we argue that one common instantiation of this procedure—evaluation with a *standard split*—is insufficient for system comparison, and propose an alternative based on multiple *random splits*.

Standard split evaluation can be formalized as follows. Let G be a set of ground truth data, partitioned into a training set G_{train}, a development set G_{dev} and a test (evaluation) set G_{test}. Let S be a system with arbitrary parameters and hyperparameters, and let $\mathcal{M}$ be an evaluation metric. Without loss of generality, we assume that $\mathcal{M}$ is a function with domain $G \times S$ and that higher values of $\mathcal{M}$ indicate better performance. Furthermore, we assume a supervised training scenario in which the free parameters of S are set so as to maximize $\mathcal{M}(G_{train}, S)$, optionally tuning hyperparameters so as to maximize $\mathcal{M}(G_{dev}, S)$. Then, if S_1 and S_2 are competing systems so trained, we prefer S_1 to S_2 if and only if $\mathcal{M}(G_{test}, S_1) > \mathcal{M}(G_{test}, S_2)$.

1.1 Hypothesis testing for system comparison

One major concern with this procedure is that it treats $\mathcal{M}(G_{test}, S_1)$ and $\mathcal{M}(G_{test}, S_2)$ as exact quantities when they are better seen as estimates of random variables corresponding to true system performance. In fact many widely used evaluation metrics, including accuracy and F-score, have known statistical distributions, allowing hypothesis testing to be used for system comparison.

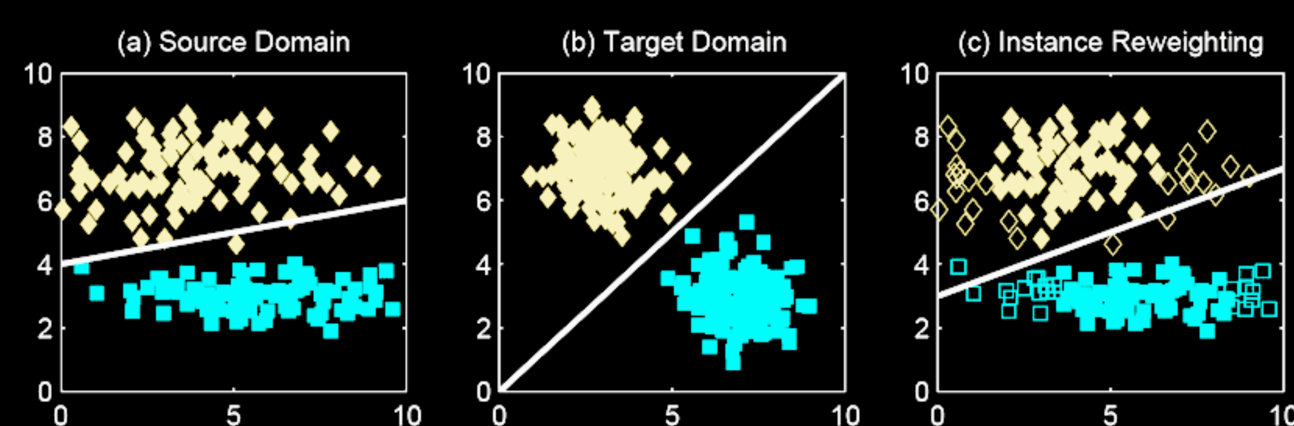
For instance, consider the comparison of two systems S_1 and S_2 trained and tuned to maximize accuracy. The difference in test accuracy, $\hat{\delta} = \mathcal{M}(G_{test}, S_1) - \mathcal{M}(G_{test}, S_2)$, can be thought of as estimate of some latent variable δ representing the true difference in system performance. While the distribution of $\hat{\delta}$ is not obvious, the probability that there is no population-level difference in system performance (i.e., $\delta = 0$) can be computed indirectly using McNemar's test (Gillick and Cox, 1989). Let $n_{1>2}$ be the number of samples in G_{test} which S_1 correctly classifies but S_2 misclassifies, and $n_{2>1}$ be the number of samples which S_1 misclassifies but S_2 correctly classifies. When $\delta = 0$, roughly half of the disagreements should favor S_1 and the other half should favor S_2. Thus, under the null hypothesis, $n_{1>2} \sim \text{Bin}(n, .5)$ where $n = n_{1>2} + n_{2>1}$. And, the (one-sided) probability of the null hypothesis is the probability of sampling $n_{1>2}$ from this distribution. Similar methods can be used for other evaluation metrics, or a reference distribution can be estimated with bootstrap resampling (Efron, 1981).

Despite this, few recent studies make use of statistical system comparison. Dror et al. (2018) survey statistical practices in all long papers presented at the 2017 meeting of the Association for Computational Linguistics (ACL), and all articles published in the 2017 volume of the *Transactions of the ACL*. They find that the majority of these works

2786

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2786–2791
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

Both reply-to and
position paper



Gorman & Bedrick (2019)

Standard splits 

Random splits 

Standard splits

Random splits

Adversarial splits

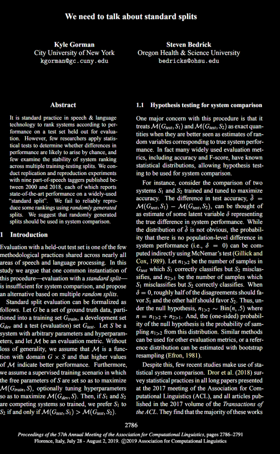
Independent samples

Standard splits

Random splits

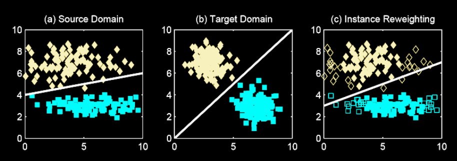
Adversarial splits

Independent samples


We need to talk about standard splits

Kyle Gorman
City University of New York

Steven Bedrick
Oregon Health & Science University



Standard splits

Gorman and Bendrick (2019): Standard splits are arbitrary and lead to community-wide overfitting.

Random splits

Random splits artificially remove sample bias and lead to overly optimistic performance estimates.

Adversarial splits

In fact, performance in real life is often worse than what can be estimated using adversarial splits.

Independent samples

It seems there's no way around evaluating our models across multiple datasets.

Core idea

Compare performance numbers across standard, random, and adversarial splits, as well as on new samples.

Probing

Headlines

POS Tagging

Core idea

Compare performance numbers across standard, random, and adversarial splits, as well as on new samples.

Quality Estimation

Emoji Prediction

News Classification

POS Tagging

Headlines

Probing

Core idea

Compare performance numbers across standard, random, and adversarial splits, as well as on new samples.

Problem

Numbers across different splits are apples and pears.

Quality Estimation

News Classification

Emoji Prediction

Comparison

Comparison

Error reduction over
(averaged) random baseline

Comparison

Error reduction over (averaged) random baseline

Say on a split, our random baseline (on average) is 0.6. SoA is 0.8. The number of interest then is…

Comparison

Error reduction over (averaged) random baseline

Say on a split, our random baseline (on average) is 0.6. SoA is 0.8. The number of interest then is…

0.33

Comparison

Error reduction over (averaged) random baseline

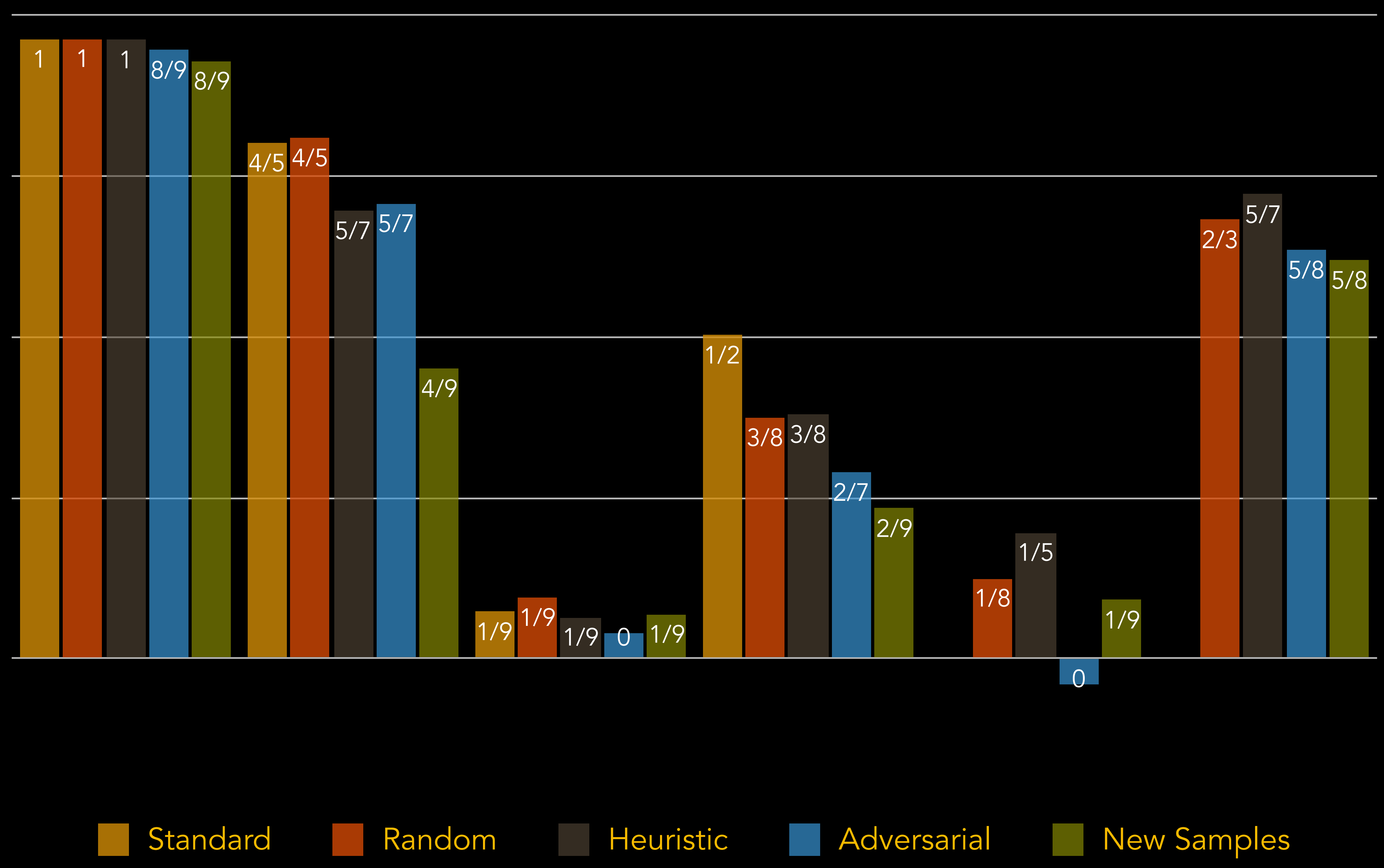
Say on a split, our random baseline (on average) is 0.6. SoA is 0.8. The number of interest then is…

0.33

Our hypothesis: These numbers are much lower on New Samples than on random splits (because overfitting).

Comparison	Error reduction over (averaged) random baseline
Heuristic	Simple heuristic, e.g., top-k longest sentences in test

Comparison	Error reduction over (averaged) random baseline
Heuristic	Simple heuristic, e.g., top-k longest sentences in test
Adversarial	Maximize Wasserstein distance



1
1
1
8/9
8/9
4/5
4/5
5/7
5/7
4/9
1/9
1/9
1/9
0
1/9
1/2
3/8
3/8
2/7
2/9
1/8
1/5
0
1/9
2/3
5/7
5/8
5/8
Standard
Random
Heuristic
Adversarial
New Samples

?

coAStaL

UNIVERSITY OF COPENHAGEN