EXPLAINABOARD: An Explainable Leaderboard for NLP

Pengfei Liu[1†], **Jinlan Fu**[2], **Yang Xiao**[2], **Weizhe Yuan**[1], **Shuaichen Chang**[3], **Junqi Dai**[2], **Yixin Liu**[1], **Zihuiwen Ye**[1], **Graham Neubig**[1‡]
[1]Carnegie Mellon University, [2]Fudan University, [3]The Ohio State University,
[†]pliu3@cs.cmu.edu, [‡]gneubig@cs.cmu.edu

Abstract

With the rapid development of NLP research, leaderboards have emerged as one tool to track the performance of various systems on various NLP tasks. They are effective in this goal to some extent, but generally present a rather simplistic one-dimensional view of the submitted systems, communicated only through holistic accuracy numbers. In this paper, we present a new conceptualization and implementation of NLP evaluation: the EXPLAINABOARD, which in addition to inheriting the functionality of the standard leaderboard, also allows researchers to (i) diagnose strengths and weaknesses of a single system (e.g. what is the best-performing system bad at?) (ii) interpret relationships between multiple systems. (e.g. where does system A outperform system B? What if we combine systems A, B, C?) and (iii) examine prediction results closely (e.g. what are common errors made by multiple systems or and in what contexts do particular errors occur?). EXPLAINABOARD has been deployed at http://explainaboard.nlpedia.ai/, and we have additionally released our interpretable evaluation code at https://github.com/neulab/ExplainaBoard along with output files[1] from more than 300 systems, 40 datasets and 9 tasks to motivate the "output-driven" research in the future.

1 Introduction

Natural language processing (NLP) research has been and is making astounding strides forward. This is true both for classical tasks such as machine translation (Sutskever et al., 2014; Wu et al., 2016), as well as for new tasks (Lu et al., 2016; Rajpurkar et al., 2016), domains (Beltagy et al., 2019), and languages (Conneau and Lample, 2019). One way this progress is quantified is through *leaderboards*, which report and update performance numbers

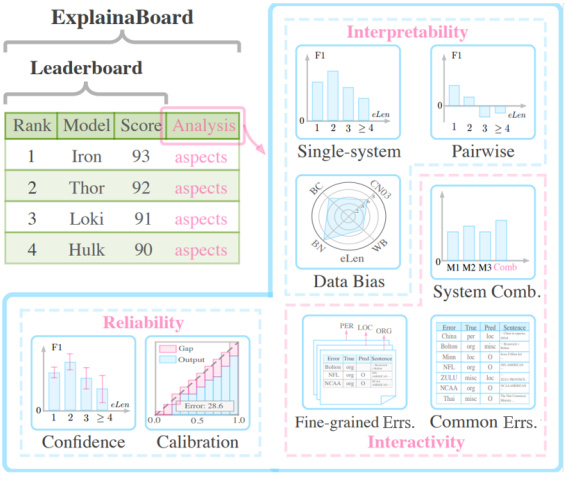

Figure 1: Illustration of the EXPLAINABOARD concept. Compared to vanilla leaderboards, EXPLAINABOARD allows users to perform *interpretable* (single-system , pairwise analysis, data bias), *interactive* (system combination, fine-grained/common error analysis), and *reliable* analysis (confidence interval, calibration) on systems in which they are interested. "comb." denotes "combination" and "Errs" represents "errors". "PER, LOC, ORG" refer to different labels.

of state-of-the-art systems on one or more tasks. Some prototypical leaderboards include GLUE and SuperGLUE for natural language understanding (Wang et al., 2018, 2019), XTREME and XGLUE (Hu et al., 2020; Liang et al., 2020) for multilingual understanding, the WMT shared tasks (Barrault et al., 2020) for machine translation, and GEM and GENIE for natural language generation (Gehrmann et al., 2021; Khashabi et al., 2021), among many others.

These leaderboards serve an important purpose: they provide a standardized evaluation setup, often on multiple tasks, that eases reproducible model comparison across organizations. However, at the same time, due to the prestige imbued by a top, or high, place on a leaderboard they also can result in a singular focus on raising evaluation numbers at the cost of deeper scientific understanding of

[1]http://explainaboard.nlpedia.ai/download.html

model properties (Ethayarajh and Jurafsky, 2020). In particular, we argue that, among others, the following are three major limitations of the existing leaderboard paradigm:

- **Interpretability:** Most existing leaderboards commonly use a single number to summarize system performance holistically. This is conducive to system ranking but at the same time the results are opaque, making the strengths and weaknesses of systems less interpretable.
- **Interactivity:** Existing leaderboards are static and non-interactive, which limits the ability of users to dig deeper into the results. Thus, (1) they usually do not flexibly support more complex evaluation settings (e.g. multi-dataset, multi-metric, multi-language) (2) users may miss opportunities to understand the relationships between different systems. For example, where does model A outperform model B? Would the performance be further improved if we combine the Top-3 models?
- **Reliability:** Given the increasing role that leaderboards have taken in guiding NLP research, it is important that information expressed in them is reliable, especially on datasets with small sample sizes, but most current leaderboards do not give an idea of the reliability of system rankings.

In this paper, we describe EXPLAINABOARD (see Fig.1), a software package and hosted leaderboard that satisfies all of the above desiderata. It also serves as a prototype implementation of some desirable features that may be included in future leaderboards, even independent of the provided software itself. We have deployed EXPLAINABOARD for 9 different tasks and 41 different datasets, and demonstrate how it can be easily adapted to new tasks of interest.

We expect that EXPLAINABOARD will benefit different steps of the research process:
(i) **System Developement**: EXPLAINABOARD provides more detailed information regarding the submitted systems (e.g. fine-grained results, confidence intervals), allowing system developers to diagnose successes and failures of their own systems, or compare their systems with baselines and understand where improvements of their proposed methods come from. This better understanding can lead to more efficient and effective system improvements. Additionally, EXPLAINABOARD can help system developers uncover their systems' advantages over others, even when these systems have not achieved state-of-the-art performance holistically. (ii) **Leaderboard Organization**: The EXPLAINABOARD software both provides a ready-made platform for easy leaderboard development over different NLP tasks, and helps upgrade traditional leaderboards to allow for more fine-grained analysis. (iii) **Broad Analysis and Understanding**: Because EXPLAINABOARD encourages system developers to provide their system outputs in an easy-to-analyze format, these will also help researchers, particularly those just starting in a particular NLP sub-field, get a broad sense of what current state-of-the-art models can and cannot do. This not only helps them quickly track the progress of different areas, but also can allow them to understand the relative advantages of diverse systems, suggesting insightful ideas for what's left and what's next.

2 ExplainaBoard

As stated above, EXPLAINABOARD extends existing leaderboards, improving their interpretability, interactivity, and reliability. It does so by providing a number of functionalities that are applicable to a wide variety of NLP tasks (as illustrated in Tab. 1). Many of these functionalities are grounded in existing research on evaluation and fine-grained diagnostics.

2.1 Interpretability

Interpretable evaluation (Popović and Ney, 2011; Stymne, 2011; Neubig et al., 2019; Fu et al., 2020a), is a research area that considers methods that break down the holistic performance of each system into different interpretable groups. For example, in a Named Entity Recognition (NER) task, we may examine the accuracy along different dimensions of a concerned entity (such as "entity frequency," telling us how well the model does on entities that appear in the training data a certain number of times) or sentences (such as "sentence length," telling us how well the model does on entities that appear in longer or shorter sentences) (Fu et al., 2020a). This makes it possible to understand where models do well and poorly, leading to further insights beyond those that can be gleaned by holistic evaluation numbers. Applying this to a new task involves the following steps: (i) *Attribute definition*: define attributes by which we can partition the test set into different groups. (ii) *Bucketing*: partition into different buckets based on defined attributes and

Aspect	Functionality	Input	Output
Interpretability	Single-system Analysis	One model	Performance Histogram: the input model is good at dealing with short entities, while achieving lower performance on long entities.
	Pairwise Analysis	Two models (M1,M2)	Performance Gap Histogram(M1 M2): M1 is better at dealing with short entities, while M2 is better at dealing with long entities.
	Data Bias Analysis	Multi-dataset	Data Bias Chart: For the entity length attribute, the average entity length (We average the length of all test entities on a given data set.) of these datasets order by descending is: BN BC> CN03> WB.
Interactivity	Fine-grained Error Analysis	Single-system or /Pairwise diagnostic results	Error Table : Error analysis allows the user to print out the entities that are incorrectly predicted by the given model, as well as the true label of the entity, the mispredicted label, and the sentence where the entity is located.
	System Combination	Multi-models (M1,M2,M3)	Ensemble Chart: The combined result of model M1, M2, and M3 is shown by the histogram with x-label value comb. The combined result is better than the single models.
Reliability	Con dence	One model	Error Bars : the error bars represent 95% con- dence intervals of the performance on the speci c bucket.
	Calibration	One model	Reliability Diagram : Con dence histograms (red) and reliability diagrams (blue). that indicate the accuracy of model probability estimates

Table 1: A graphical breakdown of the functionality of EXPLAINABOARD, with examples from an NER task.

calculate performance w.r.t each bucket.

Generally, previous work on interpretable evaluation has been performed over single tasks, while EXPLAINABOARD allows for comprehensive evaluation of different types of tasks in a single software package. We concretely show several ways interpretable evaluation can be de ned within EXPLAINABOARD below:

F1[2]: Single-system Analysis: What is a system good or bad at? For an individual system as input, generate a performance histogram that highlights the buckets where it performs well or poorly. For example, in Tab. 1 we demonstrate an example from NER where the input system does worse in dealing with longer entities (eLen 4).

F2: Pairwise Analysis: Where is one system better (worse) than another? Given a pair of systems, interpret where the performance gap occurs. Researchers could exibly choose two systems they are interested in (e.g. selecting two rows from the leaderboard), and EXPLAINABOARD will output a performance gap histogram to describe how the performance differences change over different buckets of different attributes. Tab. 1 demonstrates how we can see one system is better than the other at longer or shorter entities.

F3: Data Bias Analysis: What are the characteristics of different evaluated datasets? The de ned attributes do not only help us interpret system performance, but also make it possible for users to take a closer look at characteristics of diverse datasets. For example, from Fig. 1 shows an example of analyzing differences in average entity length across several datasets.

2.2 Interactivity

EXPLAINABOARD also allows users to dig deeper, interacting with the results in more complex ways.

[2]"F" represents "Functionality".