Square One Bias in NLP:

Towards a Multi-Dimensional Exploration of the Research Manifold

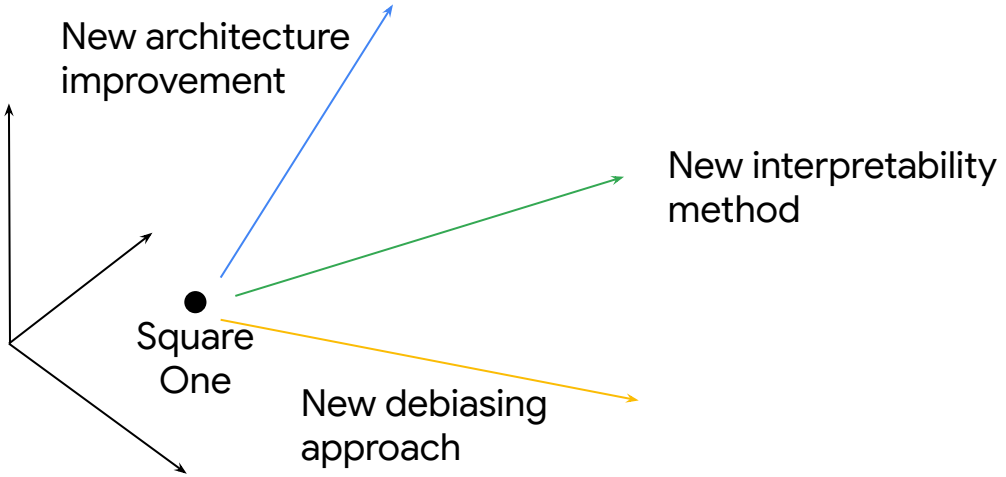
Sebastian Ruder*, Ivan Vulić*, Anders Søgaard*

Google Research

*: Equal contribution

What is the Prototypical NLP Experiment?

- Pre-trained model on a standard dataset such as SQuAD
- The existence of such an experimental prototype **steers and biases the research dynamics in our community**
- We refer to this as the **Square One** and the associated bias as **Square One Bias**
- Most research papers in NLP go beyond the prototype but *only along a single dimension*

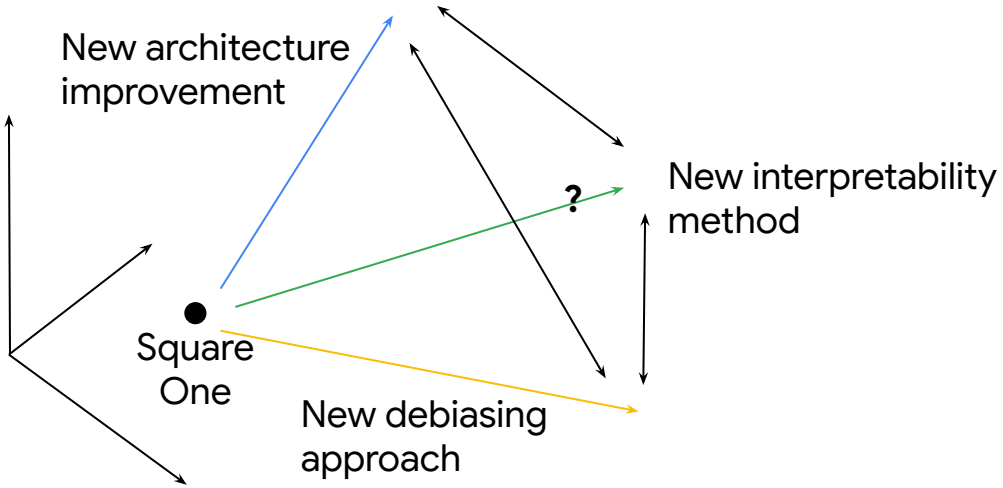


Why is this problematic?

Square One Bias

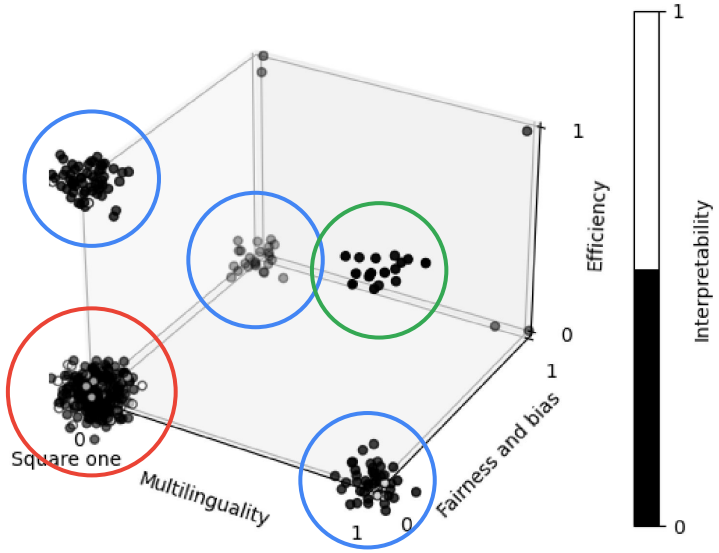
- The Square One Bias rewards work that **differs from the prototype** in a **concise, one-dimensional way**
- The study of each dimension is **biased by ignoring the others**
- By only exploring the edges of the research manifold, we are **not able to identify the non-linear interactions** between different research dimensions

→ Biases and blind spots



Finding the Square One

- Define four *common*, *general*, and *task-agnostic* dimensions of experimental contributions: **multilinguality**, **fairness and bias**, **efficiency**, and **interpretability**
- Annotate the 461 ACL 2021 oral papers based on these dimensions
- **57% of all papers** only study the standard experimental setting—evaluating **only on English** and optimizing only for a **performance metric** (such as accuracy, F1)
- **More papers** make contributions to **multilinguality** (14%) than **fairness** (6%)
- Only **6% of all papers** make a contribution along **two or more dimensions**



Visualization of contributions of ACL 2021 oral papers along 4 dimensions.

Area-level Biases

Area	# papers	English	Accuracy / F1	Multilinguality	Fairness and bias	Efficiency	Interpretability	>1 dimension
ACL 2021 oral papers	461	69.4%	38.8%	13.9%	6.3%	17.8%	11.7%	6.1%
MT and Multilinguality	58	0.0%	15.5%	56.9%	5.2%	19.0%	6.9%	13.8%
Interpretability and Analysis	18	88.9%	27.8%	5.6%	0.0%	5.6%	66.7%	5.6%
Ethics in NLP	6	83.3%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Dialog and Interactive Systems	42	90.5%	21.4%	0.0%	9.5%	23.8%	2.4%	2.4%
Machine Learning for NLP	42	66.7%	40.5%	19.0%	4.8%	50.0%	4.8%	9.5%
Information Extraction	36	80.6%	91.7%	8.3%	0.0%	25.0%	5.6%	8.3%
Resources and Evaluation	35	77.1%	42.9%	5.7%	8.6%	5.7%	14.3%	5.7%
NLP Applications	30	73.3%	43.3%	0.0%	10.0%	20.0%	10.0%	0.0%
Sentiment Analysis	18	100.0%	72.2%	0.0%	0.0%	11.1%	11.1%	0.0%
Summarization	12	91.7%	0.0%	0.0%	8.3%	0.0%	8.3%	0.0%

- Contributions along a dimension are most common within the associated area
- A large majority of work in dialog, summarization, and sentiment analysis is done only on English
- Systems in the most popular tracks evaluate on efficiency but generally do not consider fairness or interpretability
- Best papers and test-of-time papers are *not* more multi-dimensional

Examples of Square One Bias

- **Architectural**: Languages with **poor morphology**, **fixed word order**; **black-box models**; **short inputs**
- **Annotation**: Slight nuances in annotation guidelines yield model biases; interaction of annotation with multilinguality, fairness, efficiency, interpretability; translation artefacts; inflation of performance
- **Selection:** Data domain impacts properties of data items (length, demographics, style) and task requirements (fairness, efficiency)
- **Protocol:** Choice of source languages and ‘standard’ architectures
- **Organizational:** distribution of papers over well-defined areas, single-dimensional papers

Research Blind Spots

1. Trade-off between efficiency and fairness
2. Fairness and interpretability
3. Multilinguality and interpretability

Practical Recommendations

1. **New tools** and **benchmarks** to facilitate evaluation of dimensions beyond performance

2. Exploration of **'research dimension checklists'** for conference paper submission and matching of reviewers based on **complementary expertise**

3. **Awareness of research prototypes** and **prioritization** of research that departs from prototypes in **multiple dimensions**.