The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale

Keisuke Sakaguchi,* **Ronan Le Bras,*** **Chandra Bhagavatula,*** **Yejin Choi*†**
*Allen Institute for Artificial Intelligence, †University of Washington
{keisukes, ronanlb, chandrab, yejinc}@allenai.org

Abstract

The Winograd Schema Challenge (WSC) (Levesque, Davis, and Morgenstern 2011), a benchmark for commonsense reasoning, is a set of 273 expert-crafted pronoun resolution problems originally designed to be unsolvable for statistical models that rely on selectional preferences or word associations. However, recent advances in neural language models have already reached around 90% accuracy on variants of WSC. This raises an important question whether these models have truly acquired robust commonsense capabilities or whether they rely on spurious biases in the datasets that lead to an overestimation of the true capabilities of machine commonsense.

To investigate this question, we introduce WINOGRANDE, a large-scale dataset of 44k problems, inspired by the original WSC design, but adjusted to improve both the scale and the hardness of the dataset. The key steps of the dataset construction consist of (1) a carefully designed crowdsourcing procedure, followed by (2) systematic bias reduction using a novel AFLITE algorithm that generalizes human-detectable *word associations* to machine-detectable *embedding associations*. The best state-of-the-art methods on WINOGRANDE achieve 59.4 – 79.1%, which are ∼15-35% (absolute) below human performance of 94.0%, depending on the amount of the training data allowed (2% – 100% respectively).

Furthermore, we establish new state-of-the-art results on *five* related benchmarks — WSC (→ **90.1%**), DPR (→ **93.1%**), COPA(→ **90.6%**), KnowRef (→ **85.6%**), and Winogender (→ **97.1%**). These results have dual implications: on one hand, they demonstrate the effectiveness of WINOGRANDE when used as a resource for transfer learning. On the other hand, they raise a concern that we are likely to be overestimating the true capabilities of machine commonsense across all these benchmarks. We emphasize the importance of algorithmic bias reduction in existing and future benchmarks to mitigate such overestimation.

1 Introduction

The Winograd Schema Challenge (WSC) (Levesque, Davis, and Morgenstern 2011), proposed as an alternative to the Turing Test (Turing 1950), has been used as a benchmark for evaluating commonsense reasoning. WSC are designed to be pronoun resolution problems (see examples in Figure 1) that are trivial for humans but hard for machines that merely rely on statistical patterns without true capabilities of commonsense reasoning. However, recent advances in neural language models have already reported around 90% accuracy on a variant of WSC dataset.[1] This raises an important question:

> *Have neural language models successfully acquired commonsense or are we overestimating the true capabilities of machine commonsense*?

This question about the potential overestimation leads to another crucial question regarding potential unwanted biases that the large-scale neural language models might be exploiting, essentially solving the problems *right*, but for *wrong* reasons. While WSC questions are expert-crafted, recent studies have shown that they are nevertheless prone to incidental biases. Trichelair et al. (2018) have reported *word-association* (13.5% of the cases, see Figure 1 for examples) as well as other types of *dataset-specific* biases. While such biases and annotation artifacts are not apparent for individual instances, they get introduced in the dataset as problem authors subconsciously repeat similar problem-crafting strategies.

To investigate this question about the true estimation of the machine commonsense capabilities, we introduce **WINOGRANDE**, a new dataset with 44k problems that are inspired by the original design of WSC, but modified to improve both the scale and hardness of the problems. The key steps in WINOGRANDE construction consist of (1) a carefully designed crowdsourcing procedure, followed by (2) a novel algorithm AFLITE that generalizes human-detectable biases based on *word* occurrences to machine-detectable biases based on *embedding* occurrences. The key motivation of our approach is that it is difficult for humans to write problems without accidentally inserting unwanted biases.

While humans find WINOGRANDE problems trivial with 94% accuracy, best state-of-the-art results, including those from RoBERTa (Liu et al. 2019) are considerably lower,

[1] https://github.com/pytorch/fairseq/tree/master/examples/roberta. We note that this variant aggregates the original WSC, PDP (Morgenstern, Davis, and Ortiz 2016) and additional PDP-style examples, and recasts them into True/False binary problems.

		Twin sentences	Options (**answer**)
✓ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**
✓ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree** / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree / **roof**
✗ (4)	a	The lions ate the zebras because **they** are *predators*.	**lions** / zebras
	b	The lions ate the zebras because **they** are *meaty*.	lions / **zebras**

Figure 1: WSC problems are constructed as pairs (called *twin*) of nearly identical questions with two answer choices. The questions include a *trigger word* that flips the correct answer choice between the questions. Examples (1)-(3) are drawn from WSC (Levesque, Davis, and Morgenstern 2011) and (4) from DPR (Rahman and Ng 2012)). Examples marked with ✗ have language-based bias that current language models can easily detect. Example (4) is undesirable since the word "predators" is more often associated with the word "lions", compared to "zebras"

ranging between 59.4% - 79.1% depending on the amount of training data provided (from 800 to 41k instances), which falls 15 - 35% (absolute) below the human-level performance.

Furthermore, we also demonstrate that WINOGRANDE provides transfer learning to other existing WSC and related benchmarks, achieving new SOTA performances on *five* of them, including the original WSC (Levesque, Davis, and Morgenstern 2011) (→ **90.1%**), DPR (Rahman and Ng 2012) (→ **93.1%**), COPA (Roemmele, Bejan, and Gordon 2011) (→ **90.6%**), KnowRef (Emami et al. 2019) (→ **85.6%**), and Winogender (Rudinger et al. 2018) (→ **97.1%**).

Although the improvements of SOTA over multiple challenging benchmarks are exciting, we cautiously note that these positive results must be taken with a grain of salt. The result might also indicate the extent to which spurious effects are prevalent in existing datasets, which runs the risk of overestimating the true capabilities of machine intelligence on commonsense reasoning. More generally, human-crafted problems and tasks (regardless of whether they are crowdsourced or by experts) contains annotation artifacts in many cases, and algorithmic bias reduction such as AFLITE is essential to mitigate such dataset-specific bias.

2 Crowdsourcing WINOGRANDE at Scale

WSC problems have been considered challenging to craft by crowdsourcing due to the structural constraints of twins and the requirement of linguistic knowledge (Table 1). Nevertheless, we present an effective approach to creating a large-scale dataset (WINOGRANDE) of WSC problems while maintaining its original properties – i.e. trivial for humans but hard for AI systems. Our approach consists of a carefully designed crowdsourcing task followed by a novel adversarial filtering algorithm (§3) that systematically removes biases in the data.

Enhancing Crowd Creativity Creating twin sentences from scratch puts a high cognitive load on crowd workers who thereby subconsciously resort to writing pairs that are lexically and stylistically repetitive. To encourage creativity and reduce their cognitive load, we employed *creativity from constraints* (Stokes 2005) – a psychological notion which suggests that appropriate constraints can help structure and drive creativity. In practice, crowd workers are primed by a randomly chosen topic as a suggestive context (details below), while they are asked to follow precise guidelines on the structure of the curated data.

Crowdsourcing Task We collect WINOGRANDE problems via crowdsourcing on Amazon Mechanical Turk (AMT).[2] Workers are asked to write twins sentences (as shown in Table 1) that meet the requirements for WSC problems (e.g., avoiding word association, non-zero but small edit distance). To avoid repeating the same topics, workers were instructed to randomly pick an *anchor* word(s) from a randomly assigned WikiHow article[3] and to ensure that the twin sentences contain the *anchor* word. The *anchor* word does not have to be a *trigger* word, but we ensured that it is not a function word such as *the, it, he, of*. In our pilot experiments, we found that this constraint drastically improves worker's creativity and diversity of topics. Additionally, workers were instructed to keep twin sentence length in between 15 and 30 words while maintaining at least 70% word overlap between a pair of twins.[4] Following the original WSC problems, we aimed to collect twins in two different domains – (i) social commonsense: a situation involving two same gender people with contrasting attributes, emotions, social roles, etc., and (ii) physical commonsense: a context involving two physical objects with contrasting properties, usage, locations, etc. In total, we collected 77k questions (i.e., 38k twins).

Data Validation We validate each collected question through a distinct set of three crowd workers. Following a rigorous process, a question is deemed valid if (1) the majority of the three workers chooses the correct answer option, (2) they agree that the two answer options are unambiguous (one option is clearly more plausible than the other) and (3) the question cannot be answered simply by word association in which local context around the target pronoun is given (e.g., "because **it** was going so fast." (**race car** / school bus)).[5] As a

[2] Our datasets, crowdsourcing interface, and models are available at http://winogrande.allenai.org.

[3] https://www.wikihow.com/Special:Randomizer

[4] The workers met minimum qualification in AMT: 99% approval rate, $5k$ approvals. The reward was $0.4 per twin sentences.

[5] For each sentence validation, workers were paid $0.03.

result, 68% of the questions (53k) were deemed valid and we discarded the invalid questions.

While our crowdsourcing procedure addresses some amount of instance-level biases like word association, it is still possible that the constructed dataset has *dataset-specific* biases – especially after it has been scaled up. To address this challenge, we propose a method for systematic bias reduction.

3 Algorithmic Data Bias Reduction

Several recent studies (Gururangan et al. 2018; Poliak et al. 2018; Tsuchiya 2018; Niven and Kao 2019; Geva, Goldberg, and Berant 2019) have reported the presence of *annotation artifacts* in large-scale datasets. Annotation artifacts are unintentional patterns in the data that leak information about the target label in an undesired way. State-of-the-art neural models are highly effective at exploiting such artifacts to solve problems *correctly*, but for *incorrect* reasons. To tackle this persistent challenge with dataset biases, we propose AFLITE – a novel algorithm that can systematically reduce biases using state-of-the-art contextual representation of words.

Light-weight adversarial filtering Our approach builds upon the adversarial filtering (AF) algorithm proposed by Zellers et al. (2018), but makes two key improvements: (1) AFLITE is much more broadly applicable (by not requiring over generation of data instances) and (2) it is considerably more lightweight (not requiring re-training a model at each iteration of AF). Overgenerating machine text from a language model to use in test instances runs the risk of distributional bias where a discriminator can learn to distinguish between machine generated instances and human-generated ones. In addition, AF depends on training a model at each iteration, which comes at extremely high computation cost when being adversarial to a model like BERT (Devlin et al. 2018).[6]

Instead of manually identified lexical features, we adopt a dense representation of instances using their *pre-computed* neural network embeddings. In this work, we use RoBERTa (Liu et al. 2019) fine-tuned on a small subset of the dataset. Concretely, we use 6k instances (5k for training and 1k for validation) from the dataset (containing 53k instances in total) to fine-tune RoBERTa (referred to as $\text{RoBERTa}_{\text{embed}}$). We use $\text{RoBERTa}_{\text{embed}}$ to pre-compute the embeddings for the rest of the instances (47k) as the input for AFLITE. We discard the 6k instances from the final dataset.

Next, we use an ensemble of linear classifiers (logistic regressions) trained on random subsets of the data to determine whether the representation used in $\text{RoBERTa}_{\text{embed}}$ is strongly indicative of the correct answer option. If so, we discard the corresponding instances and proceed iteratively.

Algorithm 1 provides the implementation of AFLITE. The algorithm takes as input the *pre-computed* embeddings $\mathbf{X}$ and labels $\mathbf{y}$, along with the size n of the ensemble, the training size m for the classifiers in the ensemble, the size k of the filtering cutoff, and the filtering threshold τ. At each filtering phase, we train n linear classifiers on different random partitions of the data and we collect their predictions on their corresponding validation set. For each instance, we compute its *score* as the ratio of correct predictions over the total number of predictions. We rank the instances according to their score and remove the top-k instances whose score is above threshold τ. We repeat this process until we remove fewer than k instances in a filtering phase or there are fewer than m remaining instances. When applying AFLITE to WINOGRANDE, we set $m = 10,000$, $n = 64$, $k = 500$, and $\tau = 0.75$.

Algorithm 1: AFLITE

Input: dataset $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, ensemble size n, training set size m, cutoff size k, filtering threshold τ
Output: dataset $\mathcal{D}'$

$\mathcal{D}' = \mathcal{D}$
while $|\mathcal{D}'| > m$ **do**
  `// Filtering phase`
  forall $e \in \mathcal{D}'$ **do**
    Initialize the ensemble predictions $E(e) = \emptyset$
  for *iteration* $i : 1..n$ **do**
    Random partition $(\mathcal{T}_i, \mathcal{V}_i)$ of $\mathcal{D}'$ s.t. $|\mathcal{T}_i| = m$
    Train a linear classifier $\mathcal{L}$ on $\mathcal{T}_i$
    forall $e = (\mathbf{x}, y) \in \mathcal{V}_i$ **do**
      Add $\mathcal{L}(\mathbf{x})$ to $E(e)$
  forall $e = (\mathbf{x}, y) \in \mathcal{D}'$ **do**
    $score(e) = \frac{|\{p \in E(e) \text{ s.t. } p=y\}|}{|E(e)|}$
  Select the top-k elements $\mathcal{S}$ in $\mathcal{D}'$ s.t. $score(e) \geq \tau$
  $\mathcal{D}' = \mathcal{D}' \setminus \mathcal{S}$
  if $|\mathcal{S}| < k$ **then**
    break
return $\mathcal{D}'$

This approach is also reminiscent of recent work in NLP on adversarial learning (Chen and Cardie 2018; Belinkov et al. 2019; Elazar and Goldberg 2018). Belinkov et al. (2019) propose an adversarial removal technique for NLI which encourages models to learn representations that are free of hypothesis-only biases. When proposing a new benchmark, however, we cannot enforce that any future model will purposefully avoid learning spurious correlations in the data. In addition, while the hypothesis-only bias is an insightful bias in NLI, we make no assumption about the possible sources of bias in WINOGRANDE. Instead, we adopt a more proactive form of bias reduction by relying on state-of-the-art (statistical) methods to uncover undesirable dataset shortcuts.

Assessment of AFLITE We assess the impact of AFLITE relative to two baselines: random data reduction and PMI-based filtering. In random data reduction, we randomly subsample the dataset to evaluate how a decrease in dataset size affects the bias. In PMI-based filtering, we compute the difference (f) of PMIs for each twin (t) as follows:

$$f(t_1, t_2) = \sum_{w \in t_1} \text{PMI}(y = 1; w) - \sum_{w \in t_2} \text{PMI}(y = 1; w).$$

[6] AFLITE is designed for filtering instances so that the resulting dataset is less biased, whereas the original AF algorithm (Zellers et al. 2018) is designed for "generating and modifying" individual instances, such as by creating better distractors. AFLITE and AF are therefore different in their goals and difficult to compare directly.