How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG

Paul Trichelair*¹, Ali Emami*¹, Adam Trischler², Kaheer Suleman², and Jackie Chi Kit Cheung¹

¹School of Computer Science, Mila/McGill University
²Microsoft Research Montreal
{paul.trichelair, ali.emami}@mail.mcgill.ca
{adam.trischler, kasulema}@microsoft.com
jcheung@cs.mcgill.ca

Abstract

Recent studies have significantly improved the state-of-the-art on common-sense reasoning (CSR) benchmarks like the Winograd Schema Challenge (WSC) and SWAG. The question we ask in this paper is whether improved performance on these benchmarks represents genuine progress towards common-sense-enabled systems. We make case studies of both benchmarks and design protocols that clarify and qualify the results of previous work by analyzing threats to the validity of previous experimental designs. Our protocols account for several properties prevalent in common-sense benchmarks including size limitations, structural regularities, and variable instance difficulty.

1 Introduction

The proliferation of artificial-intelligence technologies that interact with human users (e.g., dialogue systems, recommendation systems, information retrieval tools) has led to renewed interest in common-sense reasoning (CSR). The progress of these technologies and the general societal reaction toward them greatly depend on advances in CSR, since systems can seem glaringly unintelligent when they lack common sense. Common sense is vital for resolving ambiguity that arises from implicit knowledge and underspecification. Consider the following sentence:

(1) The delivery truck zoomed by the school bus because it was going so fast.

Humans resolve the pronoun it to the delivery truck with no difficulty, whereas a system without common sense might be unable to distinguish the truck from the otherwise viable candidate, the school bus. The above sentence is an example from a popular binary-choice pronoun co-reference problem called the Winograd Schema Challenge (WSC) (Levesque et al., 2011), designed to directly test a machine’s grasp of common sense. What makes sentences like (1) especially challenging for machine learning approaches is that they are formulated such that simple word co-occurrence statistics cannot resolve them at a rate above chance (i.e., the delivery truck is unlikely to co-occur with going so fast much more frequently than the school bus does in large text corpora). In the same vein, a recently proposed common-sense inference task called SWAG (Zellers et al., 2018) further challenges co-occurrence-based approaches. SWAG’s problem instances comprise a partial description, along with four candidate succeeding sentences designed to be distributionally similar. Among these, one successor is the most plausible. An example SWAG instance follows.

(2) Someone is lifting the pinata. The pinata
a) drops from the swings.
b) bounces bigger than a third.
c) slumps across his shoulder back.
d) falls on the ground.

Recently, a number of systems have attained new state-of-the-art results on WSC and SWAG by querying a language model trained on a very large corpus (Trinh and Le, 2018; Radford et al., 2019; Devlin et al., 2018). The primary goal of this paper is to examine whether one can conclude from these systems’ high performance on these CSR benchmarks that they actually possess common sense. We do so by systematically examining threats to the validity of experiments involving recent CSR models.

In particular, any study aiming to show a conclusion (e.g., that a particular model can perform CSR) is subject to threats to its internal and exter-
Table 1: Examples and distribution of associative vs. non-associative WSC instances.

| Sentence Type | Examples | Proportion |
|---------------|----------|------------|
| Non-Assoc.    | Bill passed the gameboy to John because his turn was over. | 86.5% |
| Assoc.        | I’m sure that my map will show this building; it is very famous. | 13.5% |

2 Related Work

Our work presents new findings that reinforce realizations made in the community concerning the validity of a variety of different CSR tasks, most of which are in Natural Language Inference (NLI); some of these include that state-of-the-art models often do very well while being either agnostic to the premises in the task instance (which should be crucial for resolution) or by using linguistic cues that have little or nothing to do with world-knowledge or common-sense reasoning (Gururangan et al., 2018). In similar spirit, Glockner et al. (2018) create an NLI test set specifically to show the deficiencies of state-of-the-art models in inferences that require lexical and world knowledge. Alternatively, validity checks through manual investigation as in (Kalouli et al., 2017) have revealed another NLI corpus to be vulnerable to errors and model exploitation. To the best of our knowledge, our work is the first analysis performed on two very popular CSR tasks, the WSC and SWAG, that have recently garnered considerable attention in the community and on which we are beginning to see models perform relatively well (Trinh and Le, 2018; Zellers et al., 2018).

3 Validity of CSR Experiments

We now discuss the possible threats to the validity of CSR task setups in more detail.

Predictable Structure. In general, instances from both WSC and SWAG exhibit distinctive regularities. In SWAG, the counterfactual successor sentences are generated using an LSTM language model (LM), while the true successor comes from naturally occurring text. Despite recent advances in text generation, LSTM-generated responses feature stylistic patterns, such as repeated tokens, and display an overall lack of diversity (Xie, 2017). The approach SWAG introduces to minimize stylistic artifacts, adversarial filtering, is based on fooling a discriminator that classifies successors as human- or LM-generated. Nevertheless, upon inspecting the data, we found that LM-generated successors still contain repeated tokens and other signatures. A model that exploits these patterns could perform well without using
any common sense.

An example regularity found in the WSC is that instances are often composed of two clauses connected by a causal discourse connective, like because (as in (1)). This allows for simplifying assumptions (Liu et al., 2016) or schematizations (Emami et al., 2018). The issue with exploiting these structural regularities is that systems become brittle to perturbations that would not affect the judgment of a human.

**Limited Size.** Comprising only 273 test instances, the main drawback of the Winograd Schema Challenge is its limited size and the absence of training and validation sets for hyperparameter tuning. As a result, achieving above random accuracy on the WSC does not necessarily correspond to capturing common sense; it could be the result of a lucky draw.\(^1\)

**Associativity.** The WSC task definition specifies that instances should not be resolvable via statistics that associate a candidate antecedent to other components of the sentence (Levesque et al., 2011). For example, in “The lions ate the zebras because they are predators” (Rahman and Ng, 2012), the pronoun they can be resolved to lions on the basis of a much stronger association of lions with predators than of zebras with predators. We will call this (flawed) type of instance **associative** (or non-Google-proof in (Levesque et al., 2011)). Although the WSC should contain no associative sentences, there was no rigorous enforcement of this constraint. We therefore sought to quantify the associative proportion. We only consider sentences to be associative if there is a clear argument for one antecedent being statistically preferred. Table 1 outlines some examples and gives the associative proportions of the WSC.\(^2\)

### 4 New Evaluation Protocols

To probe the limitations discussed above, we propose evaluation protocols for the WSC and SWAG and apply them to several state-of-the-art methods.

**WSC.** First, we augment the existing dataset by switching candidates in sentences whenever possible (i.e., whenever switching the candidates does not obscure the sentence or affect the rationale to make the resolution decision). For example:

1. **Original:** Emma did not pass the ball to Janie although she saw that she was open.
2. **Switched:** Janie did not pass the ball to Emma although she saw that she was open.

When switching the candidates Emma and Janie, the correct answer changes as well (from Emma to Janie). A system that relies on the entity itself to make a prediction produces the same answer when the candidates are switched, even though it should not. Thus, a system that correctly resolves both the original and the switched sentence can more confidently claim to reason about the full sentence, instead of exploiting a statistical quirk of the participant entities. We introduce two new metrics based on this observation: **accuracy on the switchable subset** before and after switching the candidates, and a **consistency score**. The consistency score is the percentage of predictions that change (as would be expected) after candidates in the switchable subset are switched. In total, we counted 131 switchable instances in the WSC, which accounts for 47% of the original problem set.\(^2\)

Taking special account of both the switchable and associative instances suggests the following evaluation protocol for a given model. First, we compute the accuracy on the original WSC and the accuracy on the switchable subset of the WSC before and after switching the candidates, and compute the corresponding consistency score. Next, we compute the accuracy on the associative subset. A model can be tailored to use statistical information about the entities but perform poorly when this cannot be exploited.

**SWAG.** When evaluating on SWAG, it is important to determine whether the prediction relies on an understanding of the context or on shallow patterns in the LM-generated counterfactuals. To isolate this effect, we remove the context from the problem instances, keeping only the four successors. Three of these are machine generated. Predicting the correct label thus amounts to discriminating the human-written successor from the machine-generated ones. By comparing the performance difference between a model that has access to the context versus one that does not, we can determine the extent to which the model actually relies on contextual reasoning.
Table 2: Evaluation of state-of-the-art methods on WSC using the proposed switchability metrics. The last three columns give numbers on the switchable subset only.

| Model                     | Full WSC Acc. | Unswitched Acc. | Switched Acc. | Consistency |
|---------------------------|---------------|-----------------|---------------|-------------|
| Single LM                 | 54.58%        | 54.96%          | 54.20%        | 56.49%      |
| Ensemble 10 LMs           | 61.54%        | 58.78%          | 49.62%        | 43.51%      |
| Ensemble 14 LMs           | 63.74%        | 63.36%          | 53.43%        | 44.27%      |
| GPT-2 117M Full           | 55.68%        | 54.20%          | 54.20%        | 26.72%      |
| GPT-2 117M Partial        | 61.54%        | 59.54%          | 52.67%        | 48.85%      |
| GPT-2 774M Full           | 64.47%        | 62.60%          | 54.96%        | 45.04%      |
| GPT-2 774M Partial        | 69.23%        | 67.94%          | 61.83%        | 63.35%      |
| Knowledge Hunter          | 57.14%        | 58.78%          | 58.78%        | 90.07%      |

In both Trinh and Le (2018) and Radford et al. (2019), the language model scores the two sentences obtained when replacing the pronoun by the two candidates. The sentence that is assigned a higher probability designates the chosen candidate. Probability is calculated via the chain rule, as the product of the probabilities of each word in the sentence. The knowledge hunting method, from Emami et al. (2018), is a rule-based system that uses search engines to gather evidence for the candidate resolutions without relying on the entities themselves. BERT, a pre-trained deep bidirectional representation, is fine-tuned for SWAG using a softmax over the four possible endings.

Table 3: Accuracy of state-of-the-art methods on associative and non-associative WSC instances.

| Model                     | Assoc. | Non-Assoc. |
|---------------------------|--------|------------|
| Single LM                 | 73.0%  | 51.7%      |
| Ensemble 10 LMs           | 91.9%  | 56.8%      |
| Ensemble 14 LMs           | 83.8%  | 60.6%      |
| GPT-2 117M Full           | 73.0%  | 53.0%      |
| GPT-2 117M Partial        | 78.4%  | 58.9%      |
| GPT-2 774M Full           | 81.1%  | 61.9%      |
| GPT-2 774M Partial        | 91.9%  | 65.7%      |
| Knowledge Hunter          | 50.0%  | 58.3%      |

Table 4: Evaluation of BERT on SWAG using the proposed metrics.

| Discriminator Model       | Accuracy |
|---------------------------|----------|
| Successor-only            | 70.0%    |
| Full model                | 80.9%    |

The GPT-2 774M language models, the largest available ones, show the highest accuracy on the WSC, despite a significant drop in accuracy on the switched subset. In addition, they show the highest consistency scores on the WSC. As for the knowledge hunting method, it performed relatively well on the entire WSC, and is 100% consistent by definition, since it does not utilize the entities themselves during resolution.

In Table 3, we present model performance on the associative and non-associative subsets of the
WSC. These demonstrate that LM-based methods perform very well on the associative sentences, as expected. However, their performance drops significantly on the non-associative subset, when information related to the candidates themselves does not give away the answer.

SWAG. The performance of the state-of-the-art model is shown in Table 4. We observe that the model can distinguish human and LM-generated endings with an accuracy of 70.0%. This suggests that a strong performance on SWAG can be obtained without any consideration of the context, and that the task may not be well-suited to evaluate CSR. Nevertheless, BERT performs at 10.9% above this score when it uses the full context, indicating that the model does possess some “understanding” of the described situation.

7 Conclusion

The function of common sense in AI systems is both important and difficult to address. This paper is an attempt to make experiments, namely those performed on the WSC and SWAG, more rigorous by examining threats to the validity of these experimental designs. Based on the protocols we introduce, we show that performing at state-of-the-art on these datasets does not necessarily imply strong common-sense reasoning capability. We are happy to see a rising interest in the WSC in the community, including very recent work by Ruan et al. (2019) and Sap et al. (2019), which reinforces the need for proper evaluation protocols. With the release of an increasing number of fine-grained inference tasks aimed at these abilities (Roemmele et al., 2011; Morgenstern et al., 2016; Wang et al., 2018; Rashkin et al., 2018; McCann et al., 2018), the issue of experimental validity in CSR will also become even more important.

Acknowledgements

This work is supported by the Natural Sciences and Engineering Research Council of Canada, and by the Canada CIFAR AI Chair program. The first authors are supported in part by Microsoft Research.

References

Donald T Campbell and Julian C Stanley. 1963. Experimental and quasi-experimental designs for research. Houghton Mifflin.
Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning.

Yu-Ping Ruan, Xiaodan Zhu, Zhen-Hua Ling, Zhan Shi, Quan Liu, and Si Wei. 2019. Exploring un-supervised pretraining and sentence structure modelling for winograd schema challenge. arXiv preprint arXiv:1904.09705.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. arXiv preprint arXiv:1904.09728.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. arXiv preprint arXiv:1806.02847.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Ziang Xie. 2017. Neural text generation: A practical guide. arXiv preprint arXiv:1711.09534.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. arXiv preprint arXiv:1808.05326.