# On the Evaluation of Common-Sense Reasoning in Natural Language Understanding

**Paul Trichelair***, **Ali Emami***
Mila, McGill University
{paul.trichelair,ali.emami}@mail.mcgill.ca

**Jackie C. K. Cheung**
Mila, McGill University
jcheung@cs.mcgill.ca

**Adam Trischler, Kaheer Suleman, Fernando Diaz**
Microsoft Research Montreal
{adam.trischler,kasulema,fernando.diaz}@microsoft.com

## Abstract

The NLP and ML communities have long been interested in developing models capable of common-sense reasoning, and recent works have significantly improved the state of the art on benchmarks like the Winograd Schema Challenge (WSC). Despite these advances, the complexity of tasks designed to test common-sense reasoning remains under-analyzed. In this paper, we make a case study of the Winograd Schema Challenge and, based on two new measures of instance-level complexity, design a protocol that both clarifies and qualifies the results of previous work. Our protocol accounts for the WSC's limited size and variable instance difficulty, properties common to other common-sense benchmarks. Accounting for these properties when assessing model results may prevent unjustified conclusions.

## 1 Introduction

There is renewed interest in common-sense reasoning given the proliferation of artificial-intelligence technologies (e.g., dialogue systems, recommendation systems, information retrieval tools). The progress of these technologies, and the general societal reaction toward them, greatly depends on advances in common-sense reasoning; systems can seem glaringly *un*intelligent when they lack common sense. Common sense is vital, for example, in natural language understanding, where it is often required to resolve ambiguity arising from implicit knowledge and under-specification. Consider the following sentence:

(1) The delivery truck zoomed by the school bus because <u>it</u> was going so **fast**.

Humans resolve the pronoun *it* to *the delivery truck* with no difficulty, whereas a system without common sense would 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) [3], 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 to be robust to statistics of word co-occurence (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).

Statements like those in the WSC occur in natural settings and in more general NLP benchmarks; for example, in standard co-reference tasks [2]. However, in broader corpora, problem instances

*equal contribution

Preprint. Work in progress.

that require common sense may be rare enough not to significantly degrade the performance of statistical systems evaluated in aggregate. Common-sense reasoning that addresses these instances rather helps to cover the tail of the data, resolving rare but glaring errors in downstream tasks. The natural direction for research is to develop specialized, common-sense-focused inference tasks and use them to train and evaluate machine learning systems.

Unfortunately, it is difficult and expensive to acquire high-quality datasets for specialized inference tasks, hence existing common-sense benchmarks are very small. For example, the WSC comprises only 273 test instances, while attempts to expand it have failed to produce datasets with the same challenging characteristics (e.g., the extended WSC [6]). Size and other limitations present steep challenges in accurately evaluating systems designed for common sense. In this work, we explore some of these limitations as manifest in the WSC. We then make proposals to orient future common-sense research. In particular, we show how to partly alleviate the size issue by augmenting the dataset through problem-specific insights and then evaluating systems on finer-grained, augmented subsets. We subdivide the data according to two properties: *switchability* and *associativity*. A WSC instance is switchable when switching the occurence of the candidates does not affect the rationale used to resolve the target pronoun, and it is associative if there is a stronger word-level association between components of the instance and just one of the candidates (see §2 and §3 for more detail). We define a new evaluation protocol that incorporates these insights and apply it to existing statistical and rule-based methods. Our results indicate that the current state-of-the-art statistical method does not achieve superior performance when the dataset is augmented and subdivided with our switching scheme, and in fact mainly exploits a small subset of highly associative problem instances.

## 2 Inherent Limits of the Winograd Schema Challenge

We now analyze and discuss the inherent limits of the Winograd Schema Challenge. Our proposed evaluation protocol accounts for these limits and may be used to help researchers avoid unjustified conclusions when using the dataset.

**Limited Size** Comprising only 273 test instances,[2] the main drawback of the Winograd Schema Challenge is its limited size and the absence of training and validation sets for (hyper)parameter tuning. For its size, it turns out that if one were to choose from a set of 10 random, binary classifiers, the best based on its performance on the WSC, there is more than a 1-in-3 chance of scoring above 55% accuracy with this chosen classifier.[3] 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. As of today, state-of-the-art accuracy on the WSC for single model performance is around 55% [8, 1]. Concluding that these models capture patterns of common-sense reasoning at this performance level could be unjustified.

**Associativity** One of the specifications for sentences in the WSC is that they should not be resolvable via statistics that associate a candidate antecedent to other components of the sentence [3]. As a negative example, in the statement "The lions ate the zebras because they are predators" [6], 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 WSC instance *associative* (also called non-*Google-proof* in [3]). Associative instances can generally be resolved based on statistical tests on text corpora. Although the WSC as originally specified should contain no associative sentences, there was no rigorous enforcement of this constraint. We therefore sought to quantify its associative proportion.[4] In our human study for this purpose, we consider a sentence associative if one antecedent immediately exhibits a stronger association with the pronoun-containing clause than the other. For example, for the WSC instance, "In the storm, the tree fell down and crashed through the roof of my house. Now I have to get it repaired.", the candidate *roof* is more obviously associated with the the clause *I have to get it repaired* than *tree* is. We only consider sentences to be associative if there is a clear argument for one antecedent being preferred. Table 1 outlines some further examples and gives the associative and non-associative proportions of the WSC (13.5% and 86.5%, respectively).

[2]Recently, 13 new sentences have been added.

[3]The mathematical justification for this is included in the extra material.

[4]The dataset with the corresponding labels are available at `https://github.com/ptrichel/evaluation-common-sense` (the details are included in the appendix).

| Sentence Type | Examples | Proportion |
|---|---|---|
| Non-Associative | The sack of potatoes had been placed above the bag of flour, so it had to be moved first.<br>Bill passed the gameboy to John because his turn was over. | 86.5% |
| Associative | I'm sure that my map will show this **building**; it is very **famous**.<br>Sam broke both his ankles and he's walking with **crutches**. But a month or so from now they should be **unnecessary**. | 13.5% |

Table 1: Examples and distribution of Associative vs. Non-associative WSC instances.

**Predictable Structure** There are, generally speaking, distinctive regularities among the WSC instances. One such regularity is that, for a high number of instances, the "special" word (the hinge word that changes the correct answer when altered) is the last word, or nearly the last word. Systems can leverage this tendency in various ways, for instance via direct attention only on the latter half of an instance [8]. Many WSC instances are composed of two clauses connected by a causal discourse connective like *because* (as in (1)), which allows for simplifying assumptions [4] or schematizations [1]. The issue with exploiting these structural regularities is that systems become brittle to perturbations that would not affect the judgment of a human.

## 3 A New Evaluation Protocol

To address the limitations discussed above, we propose a new evaluation protocol for the WSC and apply it to several state-of-the-art methods. 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). An example of such a sentence is the following:

(2) **Original sentence** *Emma* did not pass the ball to *Janie* although she saw that she was open.

(3) **Switched sentence** *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 be said more certainly 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 (correctly) after candidates in the switchable subset are switched.[5] In total, we counted 131 switchable instances in the WSC, which accounts for 47% of the original problem set.

Taking special account of both the switchable and the associative instances suggests the following evaluation protocol for a given model:

1. Compute the accuracy on the original WSC.
2. Compute the accuracy on the *switchable* subset of the WSC before and after switching the candidates, and compute the corresponding consistency score. This better characterizes the model's use of context.
3. Compute the accuracy on the *associative* subset. A model can be tailored to use statistical information about the entities themselves but perform poorly when this cannot be exploited. Performance on the WSC can be interpreted in more detail using this subset.

To demonstrate the usefulness of our new protocol, we apply it on several recently proposed systems for the WSC: specially-trained, ensembled language models (LMs) [8] and a knowledge hunting

[5]The switched version of the WSC is available at `https://github.com/ptrichel/evaluation-common-sense` (the details of the study are in the appendix).

| Model | Full WSC Acc. | Unswitched Acc. | Switched Acc. | Consistency |
|---|---|---|---|---|
| Single LM [8] | 54.58% | 54.96% | 54.20% | 56.49% |
| Ensemble 10 LMs [8] | 61.54% | 58.78% | 49.62% | 43.51% |
| Ensemble 14 LMs [8] | 63.74% | 63.36% | 53.43% | 44.27% |
| Knowledge Hunter [1] | 57.14% [6] | 58.78%[6] | 58.78%[6] | 90.07%[7] |

Table 2: Evaluation of state-of-the-art methods using the proposed switchability metrics. The last three columns give numbers on the switchable subset only.

method [1] [6]. In [8], the language model scores the two sentences obtained when replacing the pronoun by the two candidates. The sentence that is assigned higher probability under the model designates the chosen candidate. Probability is calculated via the chain rule, as the product of the probabilities assigned to each word in the sentence. The knowledge hunting method is a rule-based system that uses search engines to gather evidence for the candidate resolutions without relying on the entities themselves [1].

Performance of the state-of-the-art methods with respect to our proposed switchability metrics is shown in Table 2. We observe that accuracy is stable across the different subsets for the single LM. However, the performance of the ensembled LMs, which is initially state-of-the-art by a significant margin, falls back to near random on the switched subset. This correlates with a lower consistency score than the single LM and suggests that the two ensembles overfit to the dataset. As for the Knowledge Hunter, 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. On the other hand, the Knowledge Hunter performs best among all models on the non-associative subset but struggles with associative instances.

| Model | Associative | Non-Associative |
|---|---|---|
| Single LM [8] | 73.0% | 51.7% |
| Ensemble 10 LMs [8] | 91.9% | 56.8% |
| Ensemble 14 LMs [8] | 83.8% | 60.6% |
| Knowledge Hunter [1] | 50.0%[6] | 58.3%[6] |

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

Our evaluation protocol provides a new perspective on state-of-the-art methods for common-sense reasoning. At first glance, ensembling LMs appears to be the best strategy. The deeper analysis developed in this paper suggests that this method overfits to the dataset and relies heavily on simpler word associations. Therefore, we argue that the most promising systems for the WSC are the single LM, since this method performs well on the associative sentences without compromising generality, and the knowledge hunting approach, which performs best on the non-associative sentences and is immune to switching. Nevertheless, much room for progress on the WSC still remains.

## 4 Discussion

The function of common sense is both important and difficult to address. This paper is an attempt to make experiments, namely those performed on the Winograd Schema Challenge, more rigorous. Based on the protocol we introduce, we show that performing at a state-of-the-art level on the WSC

[6] We reproduced the results for these systems using the authors' released code available on Github; Language Model: https://github.com/tensorflow/models/tree/master/research/lm_commonsense, Knowledge Hunter: https://github.com/aemami1/Wino-Knowledge-Hunter.

[6] This is the expected accuracy. For those instances that the knowledge hunter did not have enough evidence to generate a prediction, we expect half of them to be correct by chance.

[7] This is the expected consistency. For the instances that the knowledge hunter received evidence, it maintains 100% consistency during switching. However, for the instances without evidence, we expect half of the resolution decisions to flip (out of randomness) during switching.

does not necessarily imply strong common-sense reasoning. With the release of an increasing number of fine-grained inference tasks aimed at these abilities [7, 5, 9], we hope to set a precedent for future work and emphasize the importance of analysis for future empirical studies. Designing experiments that test common-sense reasoning in machines is challenging. It is especially important in this setting, and in complex natural language tasks generally, to measure what we think or claim to be measuring.

## References


[1] A. Emami, N. De La Cruz, A. Trischler, K. Suleman, and J. C. K. Cheung. A knowledge hunting framework for common sense reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, 2018.

[2] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*, 2017.

[3] H. J. Levesque, E. Davis, and L. Morgenstern. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47, 2011.

[4] Q. Liu, H. Jiang, A. Evdokimov, Z.-H. Ling, X. Zhu, S. Wei, and Y. Hu. Probabilistic reasoning via deep learning: Neural association models. *arXiv preprint arXiv:1603.07704*, 2016.

[5] L. Morgenstern, E. Davis, and C. L. Ortiz Jr. Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, 37(1):50–54, 2016.

[6] A. Rahman and V. Ng. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789. Association for Computational Linguistics, 2012.

[7] M. Roemmele, C. A. Bejan, and A. S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95, 2011.

[8] T. H. Trinh and Q. V. Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

[9] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.

# A Dataset construction

## A.1 Switching candidates

This dataset contains the original WSC with the switched version of each sentence whenever the process does not obscure the sentence or affect the rationale used to resolve the target pronoun. To construct this dataset, we first automatically switch the two candidates.

(4) **Original sentence** *Emma* did not pass the ball to *Janie* although she saw that she was open.

(5) **Switched sentence** *Janie* did not pass the ball to *Emma* although she saw that she was open.

This process can make a sentence obscure, as in the following example:

(6) **Original sentence** Sam broke both his *ankles* and he's walking with *crutches*. But a month or so from now they should be better.

(7) **Switched sentence** Sam broke both his *crutches* and he's walking with *ankles*. But a month or so from now they should be better.

The sentence obtained is not correct as *walking with ankles* is neither semantically correct nor requires the same resolution rationale. To filter out these sentences, we asked three English native speakers, who did not have prior knowledge on the WSC, to classify the sentences as *Switchable* or *Not Switchable*. We keep the switched version of the sentence if the three annotators agreed. This procedure produces a dataset of 131 switched sentences with a high agreement as shown in Table 4.

| Statistic used | Score Switchability | Score Associativity |
|---|---|---|
| Fleiss' Kappa | 0.96 | 0.79 |

Table 4: Inter-rater agreement measured using Fleiss's Kappa for both the switching and the associativity annotations

## A.2 *Associativity*

This dataset contains the original WSC sentences labeled as *associative* or *non-associative*. Associative Winograd sentences are those in which one candidate antecedent associates strongly with the clause containing the pronoun, while the other candidate antecedent exhibits no such association strength. For example:

(8) In the storm, *the tree* fell down and crashed through *the roof* of my house. Now, I have to get [it] repaired.

Here, *the roof* can be argued to be much more strongly associated with *repaired*, and on this basis, can be used to resolve the pronoun.

An example of a non-associative sentence is:

(9) Everyone really loved *the oatmeal cookies*; only a few people liked *the chocolate chip cookies*. Next time, we should make more of [them] .

Here, we don't expect, at least *a priori*, that *oatmeal cookies* associate more than *the chocolate chip cookies* with the clause, *"we should make more of them"* and therefore can be argued to be much more robust to techniques that rely on co-occurence statistics.

We split the WSC into smaller associative and non-associatve datasets by conducting a human study similar to that in A.1. The three annotators only had access to the clause containing the pronoun (e.g. *get [it] repaired* and *Next time, we should make more of [them]* for (5) and (6) respectively), and the two candidate antecedents. Using these, they were asked to categorize a sentence as associative or non-associative according to whether or not they saw a strong association between one entity and the clause, and no such association with the other entity. We chose to consider a sentence as *associative* if the three annotators unanimously agreed. This process lead to a high inter-annotator agreement as shown in Table 4 and resulted in an *associative* dataset with 37 sentences and a *non-associative* dataset with 252 sentences (there were 42 sentences for which there was not a full agreement).

## B Lucky draw

We consider a random classifier so that for each sentence, it chooses one of the two candidates. Since the dataset is balanced, the probability of getting the correct answer is 50%. When classifying the 273 instances, the number of correct answers $X$ is a binomial random variable. The probability of getting more than 55% accuracy (more than 150 correct answers) is given by:

$$\begin{aligned}
P(X > 150) &= 1 - P(X \leq 150) \\
P(X > 150) &= 1 - \sum_{i=0}^{150} P(X = i) \\
P(X > 150) &= 1 - \sum_{i=0}^{150} \binom{273}{i} 0.5^i (1 - 0.5)^{273-i} \\
P(X > 150) &= 1 - 0.5^{273} \sum_{i=0}^{150} \binom{273}{i} \\
P(X > 150) &= 0.04
\end{aligned}$$

It shows that the probability of scoring more than 55% on the WSC using a random classifier is 4%. When repeating the experiments 10 times, the probability that one of the experiments gives an accuracy greater than 55% corresponds to $1 - P(X \leq 150)^{10} = 0.37$. Practically, on the WSC, this means that if we have a pool of 10 random classifiers, there is more than a 1-in-3 chance that one of them scores more than 55%.