Balanced COPA: Countering Superficial Cues in Causal Reasoning

Pride Kavumba[1,*] Naoya Inoue[1,2,*] Benjamin Heinzerling[2,1]
Keshav Singh[1] Paul Reisert[2,1] Kentaro Inui[1,2]
[1]Tohoku University [2]RIKEN Center for Advanced Intelligence Project (AIP)
{pkavumba, naoya-i, keshav.singh29, inui} @ecei.tohoku.ac.jp
{benjamin.heinzerling, paul.reisert} @riken.jp

1 Introduction

Pretrained language models such as ELMo [15], BERT [1], RoBERTa [12] and ALBERT [9] have led to improved performance in benchmarks of natural language understanding, in tasks such as natural language inference [NLI, 11], argumentation [14], and commonsense reasoning [10, 18]. However, recent work has identified *superficial cues* in benchmark datasets which are predictive of the correct answer, such as unbalanced token distributions and lexical overlap. Once these cues are neutralized, models perform poorly, suggesting that their good performance is an instance of the Clever Hans effect[1] [16]: Models trained on datasets with superficial cues learn heuristics for exploiting these cues, but do not develop any deeper understanding of the task.

While superficial cues have been identified in, among others, datasets for NLI [5, 13], machine reading comprehension [20], and argumentation [14], one of the main benchmarks for commonsense reasoning, namely the Choice of Plausible Alternatives [COPA, 17], has not been analyzed so far. Here we present an analysis of superficial cues in COPA.

Given a premise, such as *The man broke his toe*, COPA requires choosing the more plausible, causally related alternative, in this case either: because *He got a hole in his sock* (wrong) or because *He dropped a hammer on his foot* (correct). Our analysis reveals that COPA contains superfcial cues (§2) and that finetuned BERT [1] performs well (83.9 percent accuracy) on *easy* instances containing superficial cues, but worse (71.9 percent) on *hard* instances without such simple cues (§4.3).

To prevent models from exploiting superficial cues in COPA, we introduce *Balanced COPA* (B-COPA). B-COPA contains one additional, *mirrored* instance for each original training instance. This mirrored instance uses the same alternatives as the corresponding original instance,

The woman hummed to herself. What was the *cause* for this?

✓ She was in a good mood.

✗ She was nervous.

(a) Original COPA instance.

The woman trembled. What was the *cause* for this?

✗ She was in a good mood.

✓ She was nervous.

(b) Mirrored COPA instance.

Figure 1: A COPA instance (a) with premise and correct (✓) and wrong (✗) alternatives. Our analysis reveals that the unigram *a* (highlighted orange) is a superficial cue exploited by BERT. We neutralize such superficial cues by creating a mirrored instance (b). After mirroring, the highlighted superficial cue becomes ineffective in predicting the correct answer, since it occurs with equal probability in correct and wrong alternatives.

but introduces a new premise which matches the *wrong* alternative of the original instance, e.g. *The man hid his feet*, for which the correct alternative is now because *He got a hole in his sock* (See another example in Figure 1). Since each alternative occurs exactly once as correct answer and exactly once as wrong answer in B-COPA, the lexical distribution between correct and wrong answers is perfectly balanced. Hence, superficial cues in the original alternatives have become uninformative. B-COPA allows us to study the impact of the presence or absence of superficial cues on model performance.

In summary, our contributions are:

- We identify superficial cues in COPA that allow models to use simple heuristics instead of learning the task (§2);
- We introduce Balanced COPA, which prevents models from exploiting these cues (§3). Balanced COPA is available at http://balanced-copa.github.io;
- Comparing models on original and Balanced COPA, we find that BERT heavily exploits cues when they are present, but is also able to learn the task when they are

* Equal contribution.

[1]Named after the eponymous horse which appeared to be capable of simple mental tasks but actually relied on cues given involuntarily by its handler.

not (§4); and

- We show that ALBERT and RoBERTa do not appear to exploit superficial cues.

2 Superficial Cues in COPA

COPA requires classifying sentence pairs consisting of the first sentence, the *premise*, and a second sentence that is either cause of, effect of, or unrelated to the premise. Figure 1a shows an example of a COPA instance.

Recent work found that the strong performance of BERT and other deep neural models in benchmarks of natural language understanding can be partly or in some cases entirely explained by their capability to exploit superficial cues present in benchmark datasets. Does COPA contain such cues, as well?

One of the simplest types of superficial cues are unbalanced token distributions, i.e. tokens appearing more often or less frequently with one particular instance label than with other labels. For example, Niven and Kao [14] found that the token *not* occurs more often in one type of instance in an argumentation dataset [6].

Similarly, we identify superficial cues — in this case a single token that appears more frequently in correct alternatives or wrong alternatives — in the COPA training set. For example, *a* appears in either a correct alternative or wrong alternative in 21.2% of COPA training instances. In 57.5% of these instances, it appears in correct alternatives, 7.5% more often than expected by random chance. This suggests that a model could rely on such unbalanced distributions of tokens to predict answers based only on alternatives without understanding the task.

To test this hypothesis, we perform a dataset ablation, providing only the two alternatives as input to RoBERTa, but not the premise, following similar ablations by Gururangan et al. [5], Niven and Kao [14]. RoBERTa trained[2] in this setting, i.e. on alternatives only, achieves a mean accuracy of 59.6% ($\pm$ 2.3). This is problematic because COPA is designed as a choice between alternatives given the premise. Without a premise given, model performance should not exceed random chance. Consequently, a result better than random chance shows that the dataset allows solving the task in a way that was not intended by its creators. To fix this problem, we create a balanced version of COPA that does not suffer from unbalanced token distributions in correct and wrong alternatives.

3 Balanced COPA

To allow evaluating models on a benchmark without superficial cues, we need to make them ineffective. Our approach is to balance the token distributions in correct alternatives and wrong alternatives in the training set. With a balanced token distribution, we hope models are able to learn patterns that are more relevant for the task, e.g. a pair of causally related events, rather than superficial cues.

3.1 Data Collection

To create the B-COPA training set, we manually mirror the original training set by modifying the premise. Taking the original training set as a starting point, we duplicate the COPA instances and modify their premises so that incorrect alternatives become correct (see Fig. 1 for an example).

This approach is similar to Niven and Kao [14], who create a balanced benchmark of the Argument Reasoning Comprehension Task by negating instances, but since simple negation is not applicable to COPA, we cannot follow their approach and instead manually formulate new premises. We employed five fluent English speakers with knowledge of NLP (See [7] for annotation guidelines) to create 500 new mirrored instances. Including the original training instances, B-COPA comprises a total of 1,000 instances and is publicly available at https://balanced-copa.github.io.

3.2 Qualitative Evaluation

To verify that the mirrored instances are of similar difficulty as the original ones, we measure human performance using Amazon Mechanical Turk (AMT). We randomly sample 100 instances from the original COPA training set and 100 mirrored instances from B-COPA, and asked qualified[3] crowdworkers to solve each instance. Per HIT, we assign three crowd workers and offer 10 cents reward.

From the collected responses, we calculate the accuracy of workers (by majority voting) and inter-annotator agreement by Fleiss' Kappa [2]. This evaluation shows that our mirrored instances are comparable in difficulty to the original ones (an accuracy of 97% and 100%, and Fleiss' Kappa of 0.798 and 0.973 for B-COPA and COPA, respectively).

4 Experiments

4.1 BERT, RoBERTa and ALBERT on COPA

In this section we analyze the performance of three recent pretrained language models on COPA: BERT, RoBERTa and ALBERT. The latter two are optimized variants of BERT and achieve better performance on the SuperGLUE benchmark [21], which includes COPA.

[2]See §4.1 for experimental setup.

[3]Master qualification with at least 10,000 HIT approvals and 99% HIT approval rate.

Model	Training data	Overall	Easy	Hard
Goodwin et al. [3]	-	61.8	64.7	60.0
Gordon et al. [4]	-	65.4	65.8	65.2
Sasaki et al. [19]	-	71.4	75.3	69.0
BERT-large-FT	B-COPA	74.5 ($\pm$ 0.7)	74.7 ($\pm$ 0.4)	**74.4** ($\pm$ 0.9)
BERT-large-FT	B-COPA (50%)	74.3 ($\pm$ 2.2)	76.8 ($\pm$ 1.9)	72.8 ($\pm$ 3.1)
BERT-large-FT	COPA	**76.5** ($\pm$ 2.7)	**83.9** ($\pm$ 4.4)	71.9 ($\pm$ 2.5)
RoBERTa-large-FT	B-COPA	**89.0** ($\pm$ 0.3)	88.9 ($\pm$ 2.1)	**89.0** ($\pm$ 0.8)
RoBERTa-large-FT	B-COPA (50%)	86.1 ($\pm$ 2.2)	87.4 ($\pm$ 1.1)	85.4 ($\pm$ 2.9)
RoBERTa-large-FT	COPA	87.7 ($\pm$ 0.9)	**91.6** ($\pm$ 1.1)	85.3 ($\pm$ 2.0)
ALBERT-xxlarge-v1-FT	B-COPA	**92.3** ($\pm$ 0.3)	93.0 ($\pm$ 1.1)	**91.9** ($\pm$ 0.6)
ALBERT-xxlarge-v1-FT	B-COPA (50%)	86.7 ($\pm$ 0.6)	86.0 ($\pm$ 0.8)	87.1 ($\pm$ 0.6)
ALBERT-xxlarge-v1-FT	COPA	92.1 ($\pm$ 0.3)	**93.9** ($\pm$ 1.2)	91.1 ($\pm$ 0.8)

Table 1: Performance of fine-tuned models on Balanced COPA. *Easy*: instances with superficial cues, *Hard*: instances without superficial cues.

We convert COPA instances as follows to make them compatible with the input format required by these models. Given a COPA instance $\langle p, a_1, a_2, q \rangle$, where p is a premise, a_i is the i-th alternative, and q is the question type (either *effect* or *cause*), we construct BERT's input differently, depending on the question type. We assume that the first sentence and the second sentence in the next sentence prediction task describe a cause and an effect, respectively. Specifically, for each i-th alternative, we define the following input function:

$$\text{input}(p, a_i) = \begin{cases} \text{“[CLS] } p \text{ [SEP] } a_i \text{ [SEP]” if } q \text{ is effect} \\ \text{“[CLS] } a_i \text{ [SEP] } p \text{ [SEP]” if } q \text{ is cause} \end{cases}$$

Part of BERT's training objective includes next sentence prediction. Given a pair of sentences, BERT predicts whether one sentence can be plausibly followed by the other. For this, BERT's input format contains two [SEP] tokens to mark the two sentences and the [CLS] token, which is used as the input representation for next sentence prediction. This part of BERT's architecture makes it a natural fit for COPA. For RoBERTa and ALBERT we encode two sentences in a single segment (e.g. “<s> p a_i </s>” and “$[SEP]$ p a_i $[SEP]$” respectively).[4]

After encoding premise-alternative with BERT or RoBERTa or ALBERT, we take the first hidden representation $\boldsymbol{z_i^0}$, i.e. the one corresponding to [CLS] or <s>, in the final model layer and pass it through a linear layer for binary classification.

For training, we minimize the cross entropy loss with the logits $[y_1; y_2]$ and fine-tune BERT, RoBERTa and ALBERT's parameters. In our experiments, we use pretrained BERT-large (uncased), RoBERTa-large and ALBERT-xxlarge-v1 [22].

4.2 Training Details

For training, we consider two configurations: (i) using the original COPA training set (§4.3), and (ii) using Balanced COPA (B-COPA) (§4.4). We randomly split the training data into training data and validation data with the ratio of 9:1. For B-COPA, we make sure that a pair of original instance and its mirrored counterpart always belong to the same split in order to ensure that a model is trained without superficial cues. For testing, we use all 500 instances from the original COPA test set.

We run each experiment three times with different random seeds and average the results. We train for 10 epochs and choose the best model based on the validation score. To reduce GPU RAM usage, we set BERT, RoBERTa and ALBERT's maximum sequence length to 32, which covers all training and test instances. For BERT and RoBERTa we use Adam [8] with warmup, weight decay of 0.01, a batch size of 4, and a gradient accumulation of 8, while for ALBERT, we use Adam with no warmup or weight decay, a batch size of 16, and a gradient accumulation of 3. We optimize hyperparameters for BERT, RoBERTa and ALBERT separately on the validation set.

4.3 Evaluation on Easy and Hard subsets

To investigate the behaviour of BERT, RoBERTa and ALBERT trained on the original COPA, which contains superficial cues, we split the test set into an *Easy* subset and a *Hard* subset. The *Easy subset* consists of instances that are correctly solved by the premise-oblivious model described in §2. To account for variation between the three runs with different random seeds, we deem an instance correctly classified only if the premise-oblivous model's prediction is correct for all three runs. This results in the *Easy* subset with 190 instances and the *Hard* subset comprising the remaining 310 instances. Such an easy/hard split follows similar splits in NLI datasets [5].

We then compare BERT, RoBERTa and ALBERT with previous models on the *Easy* and *Hard* subsets.[5] As Table 1 shows, previous models perform similarly on both

[4]For ALBERT, input format for BERT yields similar results. Here we report single segment encoding.

[5]For previous models, we use the prediction keys available on http://people.ict.usc.edu/~gordon/copa.html

subsets, with the exception of Sasaki et al. [19]. Overall BERT (76.5%), RoBERTa (87.7%) and ALBERT (82.3%) considerably outperform the best previous model (71.4%). However, BERT's improvements over previous work can be almost entirely attributed to high accuracy on the *Easy* subset: on this subset, finetuned BERT-large improves 8.6 percent over the model by Sasaki et al. [19] (83.9% vs. 75.3%), but on the *Hard* subset, the improvement is only 2.9 percent (71.9% vs. 69.0%). This indicates that BERT relies on superficial cues. The difference between accuracy on *Easy* and *Hard* is less pronounced for RoBERTa and ALBERT but still suggests some reliance on superficial cues. We speculate that superficial cues in the COPA training set prevented BERT, RoBERTa and ALBERT from focusing on task-related non-superficial cues such as causally related event pairs.

4.4 Evaluation on Balanced COPA

How will BERT, RoBERTa and ALBERT behave when there are no superficial cues in the training set? To answer this question, we now train BERT, RoBERTa and ALBERT on B-COPA and evaluate on the *Easy* and *Hard* subsets. The results are shown in Table 1. The smaller performance gap between *Easy* and *Hard* subsets indicates that training on B-COPA encourages BERT, RoBERTa and ALBERT to rely less on superficial cues. Moreover, training on B-COPA improves performance on the Hard subset, both when training with all 1,000 instances in B-COPA, and when matching the training size of the original COPA (500 instances, *B-COPA 50%*). Note that training on *B-COPA 50%* exposes the model to lexically less diverse training instances than the original COPA due to the high overlap between mirrored alternatives (see §3). These results show that once superficial cues are removed, the models are able to learn the task to a high degree.

5 Conclusions

We established that COPA, an important benchmark of commonsense reasoning, contains superficial cues, specifically single tokens predictive of the correct answer, that allow models to solve the task without actually understanding it. Our experiments suggest that BERT's good performance on COPA can be explained by its ability to exploit these superficial cues. BERT performs well on *Easy* instances with such superficial cues, and comparable to previous methods on *Hard* instances without such cues. RoBERTa and ALBERT, in contrast, represents a real improvement considerably outperforms both BERT and previous methods on *Hard* instances as well.

One important question remains unanswered at present, which we plan to explore in future work: When superficial cues are present, BERT clearly exploits these cues, but RoBERTa and ALBERT do not seem to rely on them. Why do RoBERTa and ALBERT not appear to rely on superficial cues, even when they are available?

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 19K20332 and JST CREST Grant Number JPMJCR1513.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of ACL*, pages 4171–4186, Minneapolis, Minnesota, June 2019.

[2] J. L. Fleiss. *Statistical methods for rates and proportions*. Wiley, New York, 2nd edition, 1981.

[3] T. Goodwin, B. Rink, K. Roberts, and S. Harabagiu. UTDHLT: COPACETIC system for choosing plausible alternatives. In *Proc. of SemEval*, pages 461–466, Montréal, Canada, 7-8 June 2012.

[4] A. S. Gordon, C. A. Bejan, and K. Sagae. Commonsense Causal Reasoning Using Millions of Personal Stories. In *25th Conference on Artificial Intelligence (AAAI-11)*, San Francisco, CA, 2011.

[5] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. In *Proc. of ACL*, pages 107–112, New Orleans, Louisiana, June 2018.

[6] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proc. of ACL*, pages 1930–1940, New Orleans, Louisiana, June 2018.

[7] P. Kavumba, N. Inoue, B. Heinzerling, K. Singh, P. Reisert, and K. Inui. When choosing plausible alternatives, clever hans can be clever. In *Proc. of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China, Nov. 2019.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[9] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations, 2019.

[10] Z. Li, T. Chen, and B. Van Durme. Learning to rank for plausible plausibility. In *Proc. of ACL*, pages 4818–4823, Florence, Italy, July 2019.

[11] X. Liu, P. He, W. Chen, and J. Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[13] T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proc. of ACL*, pages 3428–3448, Florence, Italy, July 2019.

[14] T. Niven and H.-Y. Kao. Probing neural network comprehension of natural language arguments. In *Proc. of ACL*, pages 4658–4664, Florence, Italy, July 2019.

[15] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proc. of ACL*, pages 2227–2237, New Orleans, Louisiana, June 2018.

[16] O. Pfungst. *Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology*. Holt, Rinehart and Winston, 1911.

[17] M. Roemmele, C. A. Bejan, and A. S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University, 2011.

[18] M. Sap, H. Rashkin, D. Chen, R. L. Bras, and Y. Choi. Socialiqa: Commonsense reasoning about social interactions. *ArXiv*, abs/1904.09728, 2019.

[19] S. Sasaki, S. Takase, N. Inoue, N. Okazaki, and K. Inui. Handling multiword expressions in causality estimation. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*, 2017.

[20] S. Sugawara, K. Inui, S. Sekine, and A. Aizawa. What makes reading comprehension questions easier? In *Proc. of EMNLP*, pages 4208–4219, Brussels, Belgium, Oct.-Nov. 2018.

[21] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*, 2019.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.