Counterfactually-Augmented SNLI Training Data Does Not Yield Better Generalization Than Unaugmented Data

William Huang
New York University
will.huang@nyu.edu

Haokun Liu
New York University
haokunliu@nyu.edu

Samuel R. Bowman
New York University
bowman@nyu.edu

Abstract

A growing body of work shows that models exploit annotation artifacts to achieve state-of-the-art performance on standard crowdsourced benchmarks—datasets collected from crowdworkers to create an evaluation task—while still failing on out-of-domain examples for the same task. Recent work has explored the use of counterfactually-augmented data—data built by minimally editing a set of seed examples to yield counterfactual labels—to augment training data associated with these benchmarks and build more robust classifiers that generalize better. However, Khashabi et al. (2020) find that this type of augmentation yields little benefit on reading comprehension tasks when controlling for dataset size and cost of collection. We build upon this work by using English natural language inference data to test model generalization and robustness and find that models trained on a counterfactually-augmented SNLI dataset do not generalize better than unaugmented datasets of similar size and that counterfactual augmentation can hurt performance, yielding models that are less robust to challenge examples. Counterfactual augmentation of natural language understanding data through standard crowdsourcing techniques does not appear to be an effective way of collecting training data and further innovation is required to make this general line of work viable.

1 Introduction

While standard crowdsourced benchmarks have helped create significant progress within natural language processing (NLP), a growing body of evidence shows the existence of exploitable annotation artifacts in these datasets (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018) and that models can use artifacts to achieve state-of-the-art performance on these benchmarks (McCoy et al., 2019; Naik et al., 2018). The existence of these artifacts makes it difficult to predict out-of-domain generalization and creates uncertainty around the abilities these tasks are designed to test.

Recent work has explored using counterfactually-augmented datasets to address annotation artifacts with the intent to build more robust classifiers (Kaushik et al., 2020; Khashabi et al., 2020). These datasets are collected by first sampling a set of seed examples and then creating new examples by minimally editing the seed examples to yield counterfactual labels. This type of data collection has been found to mitigate the presence of artifacts in SNLI (Bowman et al., 2015) and is presented as a way to "elucidate the difference that makes a difference" (Kaushik et al., 2020). Further, Khashabi et al. (2020) present this as an efficient method to collect training data yielding models that are "more robust to minor variations and generalize better" (Khashabi et al., 2020). However, they also find that unaugmented datasets yield better performance than datasets with 50-50 original-to-augmented data when controlling for training set size and annotation cost.

In our work, we further study whether training with counterfactually-augmented data collected through standard crowdsourcing methods yields models with better generalization and robustness by focusing on the domain of natural language inference (NLI): the task of inferring whether a *hypothesis* is true given a true *premise*. We train and compare RoBERTa (Liu et al., 2019) trained on three different datasets: (1) the counterfactually-augmented natural language inference (CNLI) training set of 8.3k seed and augmented SNLI examples from Kaushik et al. (2020), (2) a subsampled set of 8.3k unaugmented SNLI examples to control for size, and (3) the 1.7k CNLI seed examples originally sampled from SNLI. We then compare model performances on MNLI (Williams

Proceedings of the First Workshop on Insights from Negative Results in NLP, pages 82–87

ISBN 978-1-952148-66-8

et al., 2018)—a dataset for the same task with examples out-of-domain to SNLI—and two diagnostic sets (Naik et al., 2018; Wang et al., 2019a).

We find that RoBERTa trained on CNLI yields similar performance on out-of-domain MNLI examples when compared to the unaugmented subsampled SNLI training set and that including counterfactually-augmented examples to the CNLI seed set improves generalization. Further, we find that the improvement over seed examples correspond to an increase in n-grams from the addition of augmented examples, roughly doubling the number of 4-grams, and may be a result of improved lexical diversity from a larger training set. While we see similar trends in most of our diagnostic evaluations, we also find evidence that including augmented examples can yield worse performance than only training with seed examples.

While there is evidence of the benefits of using this type of data for model evaluation (Gardner et al., 2020), we find that using counterfactually-augmented data for training yields *less* robust models. We argue that further innovation is required to effectively crowdsource counterfactually-augmented natural language understanding (NLU) data for training more robust models with better generalization.

2 Related Work

Recent works show that several NLI benchmark datasets contain exploitable annotation artifacts. Several studies (Poliak et al., 2018; Gururangan et al., 2018; Tsuchiya, 2018) show that models trained on hypothesis-only examples manage to perform as much as 35 points higher than chance. Gururangan et al. (2018) also find negation words such as *no* or *never* are strongly associated with *contradiction* predictions. Other works (Naik et al., 2018; McCoy et al., 2019) find that models can exploit premise-hypothesis word overlap to achieve state-of-the-art performance on benchmarks by using associations of high overlap with *entailment* predictions and low overlap with *neutral* predictions.

Nie et al. (2020) use an adversarial human-and-model-in-the-loop procedure to address these concerns in Adversarial NLI (ANLI). Using a model in the loop makes ANLI inherently adversarial towards the model used, and we instead focus on naturally collected human-in-the-loop augmented data.

Kaushik et al. (2020) crowdsource counterfactually-augmented NLI examples that reduce the presence of hypothesis-only bias in SNLI by providing a set of seed examples to crowdworkers and prompting them to minimally edit either the hypothesis or premise to yield a counterfactual label. Khashabi et al. (2020) present this type of data collection as an efficient method to build training sets yielding robust models that generalize better by crowdsourcing counterfactually-augmented BoolQ examples. However, they also find that augmented datasets yield similar to worse performance when the cost of augmenting an example is no cheaper than collecting a new one and the datasets are controlled for size. We differ from Kaushik et al. (2020) by focusing on performance on out-of-domain examples and from Khashabi et al. (2020) by focusing on the task of NLI instead of reading comprehension.

Gardner et al. (2020) use contrast sets written manually by NLP researchers to evaluate models on various annotated tasks. They show that most datasets require 1-3 minutes per augmented example, taking 17-50 hours to create 1,000 examples. We differ by using crowdsourced counterfactually-augmented data and focusing on their use for training instead of evaluation.

3 Experimental Setup

We perform two experiments to study the effects of counterfactually-augmented NLI training data. All experiments use RoBERTa trained on SNLI, CNLI, or CNLI seed examples originally sampled from SNLI and compare performances on various tasks. We first compare MNLI performances to evaluate the impact on model generalization to out-of-domain data. We then use the diagnostic examples from Naik et al. (2018) and the GLUE diagnostic set (Wang et al., 2019a) to study model robustness to challenge examples.

Training Data In SNLI, Bowman et al. (2015) prompt crowdworkers with a scene description premise to collect three hypothesis sentences corresponding to *entailment*, *neutral*, and *contradiction* labels, yielding 570k English premise-hypothesis pairs. Kaushik et al. (2020) collect CNLI examples by prompting crowdworkers to minimally edit seed examples sampled from SNLI to yield counterfactual labels.

For our training data, we use a subsampled set

of 8.3k examples of SNLI, the CNLI training set of 8.3k examples, and the 1.7k CNLI seed examples sampled from SNLI that is also included in the CNLI training set. We subsample SNLI to control for the fact that CNLI only consists of 8.3k examples. We subsample five sets of 8.3k SNLI examples and report results across these five.

Out-of-Domain Set We treat MNLI as our out-of-domain NLI evaluation data. In collecting MNLI examples, Williams et al. (2018) follow a similar data collection framework while expanding the diversity of their premises by sourcing them from ten sources of freely available text, yielding 433k English premise-hypothesis pairs. The data set includes 393k training examples from five of the ten sources, 20k validation examples, and 20k test examples. The validation and test examples are split in half between *matched* and *mismatched* examples, where *matched* examples come from the same five sources as training examples and *mismatched* examples come from the remaining five sources. We report validation accuracy for the combined MNLI validation set.

Diagnostic Sets Naik et al. (2018) provide NLI diagnostic sets of automatically generated challenge examples based on MNLI. These sets are split into six categories named Antonymy, Numerical Reasoning, Word Overlap, Negation, Length Mismatch, and Spelling Error. As part of GLUE, Wang et al. (2019a) provide NLI diagnostic sets of challenge examples aimed to evaluate reasoning abilities related to four broad categories: Lexical Semantics, Predicate-Argument Structure, Logic, and Knowledge. We use these sets to test model robustness to challenge examples. We refer the reader to Naik et al. (2018) and Wang et al. (2019a) for additional details on each diagnostic set.

McCoy et al. (2019) provide similar adversarial examples, but we find them too difficult for our models, with performance consistently below 3%, so we do not report performance in detail.

Implementation Our code[1] builds on `jiant v2 alpha` (Wang et al., 2019b). All experiments use `roberta-base`. For each round of training, we perform 20 runs and randomly search the hyperparameter space of learning rate {1e-5, 2e-5, 3e-5}, batch size {32, 64}, and random seed. Given the small training set size and stability benefits from longer training found in Mosbach et al. (2020), we train each run for 20 epochs using early stopping based on the respective validation sets.

[1] https://github.com/nyu-mll/CNLI-generalization

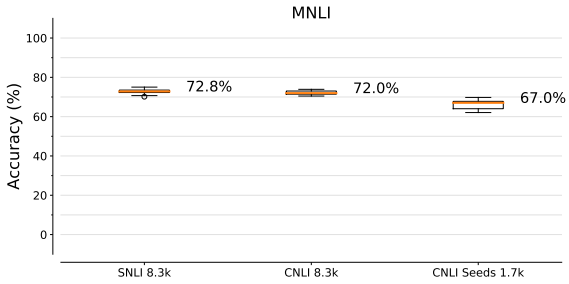


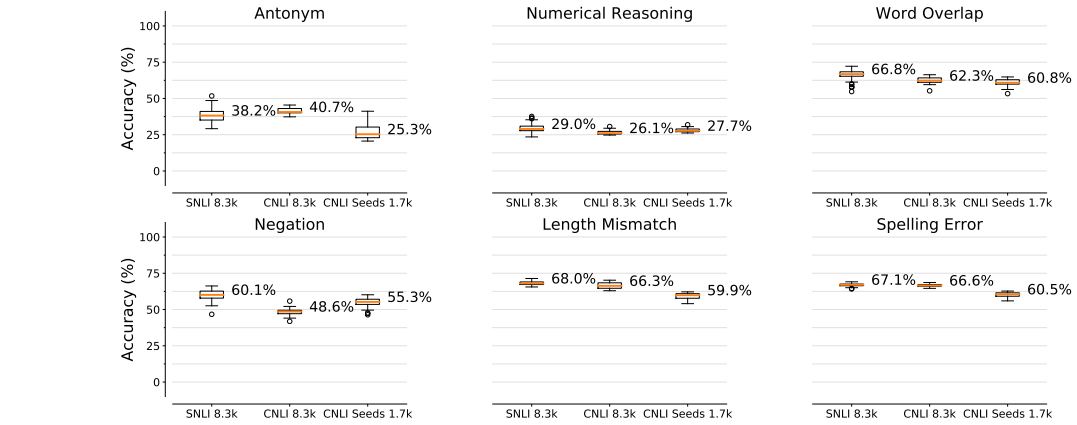
Figure 1: Combined MNLI *matched* and *mismatched* validation accuracy trained on subsampled SNLI, CNLI, and CNLI seed examples. The orange line and label indicate the median score.

4 Results

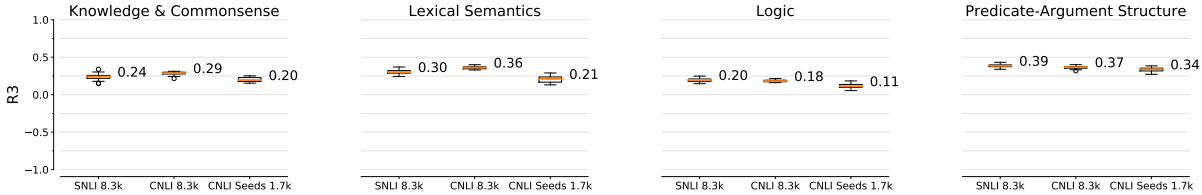
Generalization to MNLI From the median scores in Figure 1, we see that models trained on CNLI perform no better than models trained on a comparably large sample of unaugmented SNLI examples. This is in line with findings from Khashabi et al. (2020), where training with their minimally perturbed BoolQ dataset of seed and augmented examples yields similar or worse performance on out-of-domain tasks compared to the original BoolQ training set. Additionally, the improvement of CNLI over the 1.7k seed examples shows that counterfactual examples are somewhat helpful when they are strictly additive, as in Khashabi et al. (2020).

Robustness to Diagnostic Sets Figure 2 presents performances on the diagnostic sets from Naik et al. (2018) and Wang et al. (2019a). For the GLUE diagnostic sets, we follow the authors and use R_3 (Gorodkin, 2004) as our evaluation metric. The distributions of classification accuracy again show that CNLI yields similar performance compared to unaugmented datasets of similar size on most of the categories.

However, we find that training on CNLI yields worse performance than using either unaugmented SNLI or CNLI seed examples for Negation examples. These challenge examples append the phrase *"and false is not true"* to every hypothesis in the MNLI validation set. This construction introduces the strong negation word *"no"* to target the association between negation words and the *contradiction* label without changing the truth condition of the



(a)



(b)

Figure 2: Performance on diagnostic sets using (a) accuracy for Naik et al. (2018) examples and (b) R_3 score on GLUE diagnostic examples trained on subsampled SNLI, CNLI, and CNLI seed examples. Labels and orange lines indicate median scores.

n	SNLI 8.3k	CNLI 8.3k	CNLI Seed 1.7k
1	6.2k	4.8k	3.5k
2	30.4k	21.6k	13.0k
3	52.3k	36.3k	19.9k
4	60.2k	42.5k	21.5k

Table 1: Number of unique n-gram types observed in each training set.

sentence. We speculate that the augmented data may have amplified this association already present among the seed examples. Not only does this show that CNLI can yield models that are less robust to certain challenge examples, but it also provides evidence that adding substantial numbers of counterfactual examples to a dataset can hurt robustness.

Lexical Diversity Given the minimal edits constraint in CNLI, we study the lexical diversity of the training sets to see the effectiveness of this constraint and whether the general improvement of CNLI over seed examples is a result of greater diversity from a larger training set. Table 1 provides the number of n-grams present in each training set with n varying from one to four. We see that including minimally edited examples to CNLI increases the number of n-grams present, roughly doubling the number of 4-grams, which corresponds to the general improvement over seed examples.

We also observe that CNLI contains roughly 70% of 2-, 3-, and 4-grams compared to similarly large unaugmented training sets. This seems natural given the minimal edits constraint when collecting counterfactually-augmented examples and highlights the fact that this type of data augmentation results in less diversity per example.

5 Conclusion

We follow a similar setup to Khashabi et al. (2020) and use English NLI data to test whether counterfactually-augmented training data yields models that generalize better to out-of-domain data and are more robust to challenge examples. We first find that adding counterfactually-augmented data improves generalization, but provides no advantage over adding similar amounts of unaugmented data. Further, we find that the improvement over seed examples corresponds to an increase in n-gram diversity. We also find that including counterfactually-augmented data can make models less robust to challenge examples. Assuming that crowdworkers take a similar amount of time to make targeted

edits to examples and to write new examples (Bowman et al., 2020), there is then no obvious value in crowdsourcing augmentations under current protocols for use as training data.

Despite these findings, we argue that there is still value in naturally collected counterfactually-augmented NLU data. Gardner et al. (2020) show that collecting this type of data can be used as a method to address systematic gaps in testing data. As performances on benchmarks become saturated, we still view this style of augmenting test sets as a viable method to provide longer-lasting benchmarks in addition to standard test set creation.

The success of Gardner et al. (2020) in using expert-designed counterfactual augmentation to target specific phenomena for *evaluation* suggests that it may be possible to target heuristics in training data with expert guidance during the crowdsourcing process. Further, understanding how to identify heuristics to target and the types of useful augmentations to collect, assuming such a thing is possible, are important directions we leave to future work.

Acknowledgements

We thank Clara Vania and Jason Phang for their helpful feedback and Alex Wang for providing the script for n-gram counts that we base our lexical diversity analysis code on. This project has benefited from financial support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program), by Samsung Research (under the project *Improving Deep Learning using Latent Structure*), by Intuit, Inc., and in-kind support by the NYU High-Performance Computing Center and by NVIDIA Corporation (with the donation of a Titan V GPU). This material is based upon work supported by the National Science Foundation under Grant No. 1922658. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. Collecting entailment data for pretraining: New protocols and negative results. In *Proceedings of EMNLP*.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets.

J. Gorodkin. 2004. Comparing two k-category assignments by a k-category correlation coefficient. *Computational Biology and Chemistry*, 28(5):367 – 374.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of EMNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. *CoRR*, abs/2006.04884.

Aakanksha Naik, Abhilasha Ravichander, Norman M. Sadeh, Carolyn Penstein Rosé, and Graham Neubig.

2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2340–2353. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 180–191. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Phil Yeres, Jason Phang, Haokun Liu, Phu Mon Htut, , Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Edouard Grave, Najoung Kim, Thibault Févry, Berlin Chen, Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman. 2019b. `jiant` 1.3: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.