Annotation Artifacts in Natural Language Inference Data

Suchin Gururangan[★] [♢] **Swabha Swayamdipta**[★] [♡]
Omer Levy[♣] **Roy Schwartz**[♣♠] **Samuel R. Bowman** [†] **Noah A. Smith**[♣]

[♢] Department of Linguistics, University of Washington, Seattle, WA, USA
[♡] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[♣] Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[♠] Allen Institute for Artificial Intelligence, Seattle, WA, USA
[†] Center for Data Science and Department of Linguistics, New York University, New York, NY, USA
{sg01,swabha,omerlevy,roysch,nasmith}@cs.washington.edu bowman@nyu.edu

Abstract

Large-scale datasets for natural language inference are created by presenting crowd workers with a sentence (premise), and asking them to generate three new sentences (hypotheses) that it entails, contradicts, or is logically neutral with respect to. We show that, in a significant portion of such data, this protocol leaves clues that make it possible to identify the label by looking only at the hypothesis, without observing the premise. Specifically, we show that a simple text categorization model can correctly classify the hypothesis alone in about 67% of SNLI (Bowman et al., 2015) and 53% of MultiNLI (Williams et al., 2018). Our analysis reveals that specific linguistic phenomena such as negation and vagueness are highly correlated with certain inference classes. Our findings suggest that the success of natural language inference models to date has been overestimated, and that the task remains a hard open problem.

1 Introduction

Natural language inference (NLI; also known as recognizing textual entailment, or RTE) is a widely-studied task in natural language processing, to which many complex semantic tasks, such as question answering and text summarization, can be reduced (Dagan et al., 2006). Given a pair of sentences, a premise p and a hypothesis h, the goal is to determine whether or not p semantically entails h.

The problem of acquiring large amounts of labeled inference data was addressed by Bowman et al. (2015), who devised a method for crowdsourcing high-agreement entailment annotations en masse, creating the SNLI and later the genre-diverse MultiNLI (Williams et al., 2018) datasets. In this process, crowd workers are presented with a premise p drawn from some corpus (e.g., image captions), and are required to generate three new sentences (hypotheses) based on p, according to one of the following criteria:

Entailment	h is definitely true given p
Neutral	h might be true given p
Contradiction	h is definitely **not** true given p

In this paper, we observe that hypotheses generated by this crowdsourcing process contain artifacts that can help a classifier detect the correct class *without* ever observing the premise (Section 2).

A closer look suggests that the observed artifacts are a product of specific annotation strategies and heuristics that crowd workers adopt. We find, for example, that entailed hypotheses tend to contain gender-neutral references to people, purpose clauses are a sign of neutral hypotheses, and negation is correlated with contradiction (Section 3). Table 1 shows a single set of instances from SNLI that demonstrates all three phenomena.

We re-evaluate high-performing NLI models on the subset of examples on which our hypothesis-only classifier failed, which we consider to be "hard" (Section 4). Our results show that the performance of these models on the "hard" subset is dramatically lower than their performance on the rest of the instances. This suggests that, despite recently reported progress, natural language inference remains an open problem.

2 Annotation Artifacts are Common

We conjecture that the framing of the annotation task has a significant effect on the language generation choices that crowd workers make when authoring hypotheses, producing certain patterns in the data. We call these patterns *annotation artifacts*.

[★] These authors contributed equally to this work.

Proceedings of NAACL-HLT 2018, pages 107–112
New Orleans, Louisiana, June 1 - 6, 2018.

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Table 1: An instance from SNLI that illustrates the artifacts that arise from the annotation protocol. A common strategy for generating entailed hypotheses is to remove gender or number information. Neutral hypotheses are often constructed by adding a purpose clause. Negations are often introduced to generate contradictions.

Model	**SNLI**	**MultiNLI**	
		Matched	**Mismatched**
majority class	34.3	35.4	35.2
`fastText`	**67.0**	**53.9**	**52.3**

Table 2: Performance of a premise-oblivious text classifier on NLI. The MultiNLI benchmark contains two test sets: matched (in-domain examples) and mismatched (out-of-domain examples). A majority baseline is presented for reference.

To determine the degree to which such artifacts exist, we train a model to predict the label of a given hypothesis *without seeing the premise*. Specifically, we use `fastText` (Joulin et al., 2017), an off-the-shelf text classifier that models text as a bag of words and bigrams, to predict the entailment label of the hypothesis.[1] This classifier is completely oblivious to the premise.

Table 2 shows that a significant portion of each test set can be correctly classified without looking at the premise, well beyond the most-frequent-class baseline.[2]

Our finding demonstrates that it is possible to perform well on these datasets without modeling natural language inference.

3 Characteristics of Annotation Artifacts

In the previous section we showed that more than half (MultiNLI) or even two thirds (SNLI) of the data can be classified correctly using annotation artifacts. A possible explanation for the formation and relative consistency of these artifacts is that crowd workers adopt heuristics in order to generate hypotheses quickly and efficiently. We identify some of these heuristics by conducting a shallow statistical analysis of the data, focusing on lexical choice (Section 3.1) and sentence length (Section 3.2).

3.1 Lexical Choice

To see whether the use of certain words is indicative of the inference class, we compute the pointwise mutual information (PMI) between each word and class in the training set:

$$\text{PMI}(\textit{word}, \textit{class}) = \log \frac{p(\textit{word}, \textit{class})}{p(\textit{word}, \cdot)p(\cdot, \textit{class})}$$

We apply add-100 smoothing to the raw statistics; the aggressive smoothing emphasizes word-class correlations that are highly discriminative. Table 4 shows the top words affiliated with each class by PMI, along with the proportion of training sentences in each class containing each word.

Below, we elaborate on the most discriminating words for each NLI class, and suggest possible annotation heuristics that gave rise to these particular artifacts. However, it is important to note that even the most discriminative words are not very frequent, indicating that the annotation artifacts are diverse, and that crowd workers tend to adopt multiple heuristics for generating new text.

Entailment. Entailed hypotheses have generic words such as *animal*, *instrument*, and *outdoors*, which were probably chosen to generalize over more specific premise words such as *dog*, *guitar*, and *beach*. Other heuristics seem to replace exact numbers with approximates (*some*, *at least*, *various*), and to remove explicit gender (*human* and *person* appear lower down the list). Some artifacts are specific to the domain, such as *outdoors* and *outside*, which are typical of the personal photo descriptions on which SNLI was built.

[1] For MultiNLI, we additionally enabled two hyperparameters: character 4-grams, and filtering words that appeared less than 10 times in the training data.

[2] Experiments with two other text classifiers, a logistic regression classifier with word and character n-gram features and a premise-oblivious version of the decomposable attention model (Parikh et al., 2016), yielded similar results.

Premise	Two dogs are running through a field.
Entailment	There are **animals outdoors**.
Neutral	Some puppies are running **to catch a stick**.
Contradiction	The pets are **sitting on a couch**.

Table 3: The example provided in the annotation guidelines for SNLI. Some of the observed artifacts (bold) can be potentially traced back to phenomena in this specific example.

	Entailment		Neutral		Contradiction	
SNLI	outdoors	2.8%	tall	0.7%	nobody	0.1%
	least	0.2%	first	0.6%	sleeping	3.2%
	instrument	0.5%	competition	0.7%	no	1.2%
	outside	8.0%	sad	0.5%	tv	0.4%
	animal	0.7%	favorite	0.4%	cat	1.3%
MNLI	some	1.6%	also	1.4%	never	5.0%
	yes	0.1%	because	4.1%	no	7.6%
	something	0.9%	popular	0.7%	nothing	1.4%
	sometimes	0.2%	many	2.2%	any	4.1%
	various	0.1%	most	1.8%	none	0.1%

Table 4: Top 5 words by PMI($word, class$), along with the proportion of *class* training samples containing *word*. MultiNLI is abbreviated to MNLI.

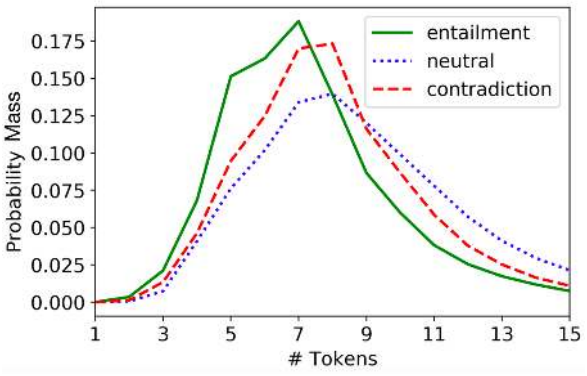

Figure 1: The probability mass function of the hypothesis length in SNLI, by class.

Interestingly, the example from the SNLI annotation guidelines (Table 3) contains both *animals* and *outdoors*, and also removes the number. This example likely primed the annotators, inducing the specific heuristics of replacing *dog* with *animal* and mentions of scenery with *outdoors*.

Neutral. Modifiers (*tall*, *sad*, *popular*) and superlatives (*first*, *favorite*, *most*) are affiliated with the neutral class. These modifiers are perhaps a product of a simple strategy for introducing information that is not obviously entailed by the premise, yet plausible. Another formulation of neutral hypotheses seems to be through cause and purpose clauses, which increase the prevalence of discourse markers such as *because*. Once again, we observe that the example from the SNLI annotation guidelines does just that, by adding the purpose clause *to catch a stick* (Table 3).

Contradiction. Negation words such as *nobody*, *no*, *never* and *nothing* are strong indicators of contradiction.[3] Other (non-negative) words appear to be part of heuristics for contradicting whatever information is displayed in the premise; *sleeping* contradicts any activity, and *naked* (further down the list) contradicts any description of clothing.

[3]Similar findings were observed in the ROC story cloze annotation (Schwartz et al., 2017).

The high frequency of *cat* probably stems from the many dog images in the original dataset.

3.2 Sentence Length

We observe that the number of tokens in generated hypotheses is not distributed equally among the different inference classes. Figure 1 shows that, in SNLI, neutral hypotheses tend to be long, while entailed ones are generally shorter. The median length of a neutral hypothesis is 9, whereas 60% of entailments have 7 tokens or less. We also observe that half of hypotheses with at least 12 tokens are neutral, while a similar portion of hypotheses of length 5 and under are entailments, making hypothesis length an effective feature. Length is also a discriminatory feature in MultiNLI, but is less significant, possibly due to the introduction of diverse genres.

The bias in sentence length may suggest that crowd workers created many entailed hypotheses by simply removing words from the premise. Indeed, when representing each sentence as a bag of words, 8.8% of entailed hypotheses in SNLI are fully contained within their premise, while only 0.2% of neutrals and contradictions exhibit the same property. MultiNLI showed similar trends.

Model	SNLI			MultiNLI Matched			MultiNLI Mismatched		
	Full	*Hard*	*Easy*	*Full*	*Hard*	*Easy*	*Full*	*Hard*	*Easy*
DAM	84.7	69.4	92.4	72.0	55.8	85.3	72.1	56.2	85.7
ESIM	85.8	71.3	92.6	74.1	59.3	86.2	73.1	58.9	85.2
DIIN	86.5	72.7	93.4	77.0	64.1	87.6	76.5	64.4	86.8

Table 5: Performance of high-performing NLI models on the full, *Hard*, and *Easy* NLI test sets.

4 Re-evaluating NLI Models

In Section 2, we showed that a model with no access to the premise can correctly classify many examples in both SNLI and MultiNLI, performing well above the most-frequent-class baseline. This raises an important question about state-of-the-art NLI models: to what extent are they "gaming" the task by learning to detect annotation artifacts?

To answer this question, we partition each NLI test set into two subsets: examples that the premise-oblivious model classified accurately are labeled *Easy*, and those it could not are *Hard*.

We then train an NLI model on the original training sets (from either SNLI or MultiNLI),[4] and evaluate on the full test set, the *Hard* test set, and the *Easy* test set. We ran this experiment on three high-performing NLI models: the Decomposable Attention Model (DAM; Parikh et al., 2016),[5] the Enhanced Sequential Inference Model (ESIM; Chen et al., 2017),[6] and the Densely Interactive Inference Network (DIIN; Gong et al., 2018).[7] All models were retrained out of the box.

Table 5 shows the performance of each model on the different splits. While the models correctly classify some *Hard* examples, the bulk of their success is attributed to the *Easy* examples. This result implies that the ability of NLI models to recognize textual entailment is lower than previously perceived, and that such models rely heavily on annotation artifacts in the hypothesis to make their predictions.

A natural question to ask is whether it is possible to select a set of NLI training and test samples which do not contain easy-to-exploit artifacts. One solution might be to filter *Easy* examples from the training set, retaining only *Hard* examples. However, initial experiments suggest that it might not be as straightforward to eliminate annotation artifacts once the dataset has been collected.

First, after removing the *Easy* examples, *Hard* examples might not necessarily be artifact-free. For instance, removing all contradicting samples containing the word "no" (a strong indicator for contradiction, see Section 3), leaves the *Hard* dataset with this word mostly appearing in the neutral and entailing classes, thus creating a new artifact. Secondly, *Easy* examples contain important inference phenomena (e.g. the word "animal" is indeed a hypernym of "dog"), and removing these examples may hinder the model from learning such phenomena. Importantly, artifacts do not render any particular example *incorrect*; they are a problem with the sample distribution, which is skewed toward certain kinds of entailment, contradiction, and neutral hypotheses. Therefore, a better solution might not eliminate the artifacts altogether, but rather balance them across labels. Future strategies for reducing annotation artifacts might involve experimenting with the prompts or training given to crowd workers, e.g., to encourage a wide range of strategies, or incorporating baseline or adversarial systems that flag examples that appear to use over-represented heuristics. We defer research on hard-to-exploit NLI datasets to future work.

5 Discussion

We reflect on our results and relate them to other work that also analyzes annotation artifacts in NLP datasets, drawing three main conclusions.

Many datasets contain annotation artifacts. Lai and Hockenmaier (2014) demonstrated that lexical features such as the presence of negation, word overlap, and hypernym relations are highly predictive of entailment classes in the SICK dataset (Marelli et al., 2014). Chen et al. (2016) revealed problems with the CNN/DailyMail dataset (Hermann et al., 2015) which resulted from apply-

[4] The MultiNLI models were trained on MultiNLI data alone (as opposed to a blend of MultiNLI and SNLI data).

[5] github.com/allenai/allennlp

[6] github.com/nyu-mll/multiNLI

[7] goo.gl/kCeZXm

ing automatic tools for annotation. Levy and Dagan (2016) showed that a relation inference benchmark (Zeichner et al., 2012) is severely biased towards distributional methods, since it was created using DIRT (Lin and Pantel, 2001). Schwartz et al. (2017) and Cai et al. (2017) showed that certain biases are prevalent in the ROC stories cloze task (Mostafazadeh et al., 2016), which allow models trained on the endings alone, and not the story prefix, to yield state-of-the-art results. Rudinger et al. (2017) revealed that elicited hypotheses in SNLI contain evidence of various gender, racial, religious, and aged-based stereotypes. In parallel to this work, Poliak et al. (2018) uncovered similar annotation biases across multiple NLI datasets. Indeed, annotation artifacts are not unique to the NLI datasets, and the danger of such biases should be carefully considered when annotating new datasets.

Supervised models leverage annotation artifacts. Levy et al. (2015) demonstrated that supervised lexical inference models rely heavily on artifacts in the datasets, particularly the tendency of some words to serve as prototypical hypernyms. Agrawal et al. (2016); Jabri et al. (2016); Goyal et al. (2017) all showed that state-of-the-art visual question answering (Antol et al., 2015) systems leverage annotation biases in the dataset. Cirik et al. (2018) find that complex models for referring expression recognition achieve high performance without any text input. In parallel to this work, Dasgupta et al. (2018) found that the InferSent model (Conneau et al., 2017) relies on word-level heuristics to achieve state-of-the-art performance on SNLI. These findings coincide with ours, and strongly suggest that supervised models will exploit shortcuts in the data for gaming the benchmark, if such exist.

Annotation artifacts inflate model performance. This is a corollary of the above, since large portions of the test set can be solved by relying on annotation artifacts alone. A similar finding by Jia and Liang (2017) showed that the performance of top question-answering models trained on SQuAD (Rajpurkar et al., 2016) drops drastically by introducing simple adversarial sentences in the evidence. We release the *Hard* SNLI and MultiNLI test sets,[8] and encourage the community to use them for evaluating NLI models (in addition to the original benchmarks). We also encourage the development of additional challenging benchmarks that expose the true performance levels of state-of-the-art NLI models.

[8] SNLI: goo.gl/5rQKb5, MultiNLI matched: goo.gl/abdSbi, MultiNLI mismatched: goo.gl/Cu9Gp6

Acknowledgments

This research was supported in part by the DARPA CwC program through ARO (W911NF-15-1-0543) and a hardware gift from NVIDIA Corporation. SB acknowledges gift support from Google and Tencent Holdings and support from Samsung Research.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proc. of EMNLP*. https://aclweb.org/anthology/D16-1203.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proc. of ICCV*. https://arxiv.org/abs/1506.00278.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proc. of EMNLP*. https://doi.org/10.18653/v1/D15-1075.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending:strong neural baselines for the ROC Story Cloze task. In *Proc. of ACL*. https://doi.org/10.18653/v1/P17-2097.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proc. of ACL*. https://doi.org/10.18653/v1/P16-1223.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, and Diana Inkpen. 2017. Natural language inference with external knowledge. arXiv:1711.04289. https://arxiv.org/abs/1711.04289.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do our systems actually learn? In *Proc. of NAACL*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proc. of EMNLP*. https://doi.org/10.18653/v1/D17-1070.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. *Machine Learning Challenges* pages 177–190. https://doi.org/10.1007/11736790_9.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. https://arxiv.org/abs/1802.04302.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In *Proc. of ICLR*. https://arxiv.org/abs/1709.04348.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proc. of CVPR*. https://arxiv.org/abs/1612.00837.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proc. of NIPS*. http://dl.acm.org/citation.cfm?id=2969239.2969428.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *Proc. of ECCV*. https://doi.org/10.1007/978-3-319-46484-8_44.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proc. of EMNLP*. https://www.aclweb.org/anthology/D17-1215.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proc. of EACL*. https://doi.org/10.18653/v1/E17-2068.

Alice Lai and Julia Hockenmaier. 2014. Illinois-LH: A denotational and distributional approach to semantics. In *Proc. of SemEval*. https://doi.org/10.3115/v1/S14-2055.

Omer Levy and Ido Dagan. 2016. Annotating relation inference in context via question answering. In *Proc. of ACL*. https://doi.org/10.18653/v1/P16-2041.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proc. of NAACL*. https://doi.org/10.3115/v1/N15-1098.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question-answering. *Natural Language Engineering* 7(4):343–360. https://doi.org/10.1017/S1351324901002765.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proc. of LREC*. pages 216–223. https://doi.org/10.3115/v1/S14-2001.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and Cloze evaluation for deeper understanding of commonsense stories. In *Proc. of NAACL*. https://doi.org/10.18653/v1/N16-1098.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proc. of EMNLP*. https://doi.org/10.18653/v1/D16-1244.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines for natural language inference. In *Proc of *SEM*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*. https://doi.org/10.18653/v1/D16-1264.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proc. of EthNLP*. http://www.aclweb.org/anthology/W17-1609.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC Story Cloze task. In *Proc. of CoNLL*. https://doi.org/10.18653/v1/K17-1004.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of NAACL*.

Naomi Zeichner, Jonathan Berant, and Ido Dagan. 2012. Crowdsourcing inference-rule evaluation. In *Proc. of ACL*. http://www.aclweb.org/anthology/P12-2031.